STATISTICAL MODELING OF PROTEIN LYSATE ARRAY DATA

BY

JI-YEON YANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Professor Xuming He, Chair
Professor Jeff Douglas
Professor Feng Liang
Professor Annie Qu

ABSTRACT

The protein lysate array is an emerging technology for quantifying the protein concentration ratios in multiple biological samples. It is gaining popularity, and has the potential to answer questions about post-translational modifications and protein pathway relationships.

Statistical inference for a parametric quantification procedure has been inadequately addressed in the literature, mainly due to two challenges: the increasing dimension of the parameter space and the need to account for dependence in the data. Each chapter of this thesis addresses one of these issues.

In Chapter 1, an introduction to the protein lysate array quantification is presented, followed by the motivations and goals for this thesis work.

In Chapter 2, we develop a multi-step procedure for the Sigmoidal models, ensuring consistent estimation of the concentration level with full asymptotic efficiency. The results obtained in this chapter justify inferential procedures based on large-sample approximations. Simulation studies and real data analysis are used to illustrate the performance of the proposed method in finite-samples. The multi-step procedure is simpler in both theory and computation than the single-step least squares method that has been used in current practice.

In Chapter 3, we introduce a new model to account for the dependence structure of the errors by a nonlinear mixed effects model. We consider a method to approximate the maximum likelihood estimator of all the parameters. Using the simulation studies on various error structures, we show that for data with *non-i.i.d.* errors the proposed method leads to more accurate estimates and better confidence intervals than the existing single-step least squares method.

To my husband, Jungmo.

ACKNOWLEDGMENTS

My first, and most sincere, acknowledgment must go to my advisor, Professor Xuming He. Without his guidance, support, and immense knowledge, this thesis work would not have been accomplished. His patience and encouragement helped me go through many difficult times. I also wish to express my appreciation to Professor Feng Liang. She was always willing to have discussions with me, and offered many valuable suggestions and encouragements. Sincere thanks are extended to other committee members, Professors Jeff Douglas and Annie Qu for their helpful discussions and insightful comments. I thank my fellow PhD students at UIUC, particularly Ji Young Kim and Ya-Hui Hsu, for their friendship and support. Lastly, I am deeply indebted to my entire family for their unconditional love, support, and dedication.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Overview

This dissertation concerns the statistical modeling of the protein lysate array data. The protein array quantification poses two challenges: the increasing dimension of the parameter space and the need to account for dependence in the data. Each chapter of this thesis addresses one of these issues. An introduction to the protein lysate array quantification along with the experimental setup for the array is presented in this chapter, followed by the motivations and goals for this thesis work.

1.2 Protein Lysate Arrays

It is known that the study of cells at the protein level is much more complex than at the genome level. However, the genomic studies alone cannot explain protein structures and do not provide insight into post-translational modifications, such as phosphorylation, acetylation and ubiquitination. A comprehensive study of both genes and proteins is necessary to understand the cellular basis of disease onset and progression ([1], [2], [3], [4], [5]). A major goal of proteomics is to identify protein changes associated with the development of diseases such as cancer. The identified proteins are potential biomarkers for the disease. Protein arrays are gaining popularity, and have the potential to answer questions about post-translational modifications and protein pathway relationships.

The protein microarray formats are mainly divided into two types: forward-phase array (FPA) and reverse-phase array (RPA). In the forward-phase array, various antibodies are robotically spotted on the slides and each slide is incubated with a sample. With this format, we can perform the simultaneous measurement of multiple proteins across a single sample ([6], [7]). The reverse-phase array has the opposite configuration. Multiple samples

are robotically spotted on the slides and each slide is incubated with an antibody that is specific for the protein of interest. As a result, the reverse-phase array, also called the protein lysate array, measures the levels of a common protein across multiple samples ([8], [9]). This thesis focuses on the second type, the protein lysate array, and aims to measure the relative level of a common protein in several samples .

In the protein lysate array, each sample that is a solution containing the target protein is serially diluted by a certain factor several times. Suppose that there are s samples assessed and each sample is 2-fold serially diluted t times. Let c_i be the protein concentration of the i^{th} sample. Then a dilution series, $c_i, c_i/2, c_i/2^2, ..., c_i/2^{(t-1)}$ is obtained from the i^{th} sample. Each dilution is spotted on a nitrocellulose-coated slide in r replicates and probed with an antibody that recognizes the target protein. Then srt spots yield gray-level intensities with higher intensity reflecting a higher concentration level of the protein. Note that the dilution and the replicates give rt intensities for each sample allowing for a more accurate measurement than with individual spot intensity. Figure 1.1 illustrates the protein lysate array with $s = 96$ biological samples, $r = 3$ replicates, and $t = 6$ dilution levels. This figure is taken from [10]. Panel (a) is the image of the protein lysate array slide for the 96 samples. Panels (b) and (c) are the magnifications corresponding to eight samples and two samples, respectively. Each sample has 18 spots with 3 replicates and 6 dilution levels. The details on the experimental setup for the protein lysate array can be found in [10].

The gray-level on each spot has a positive relationship with the protein concentration level that depends on the dilutions and we may assume the following model,

$$y_{ijl} = g(c_i/2^l) + \epsilon_{ijl},$$

where $i = 1, 2, ..., s,\ j = 1, 2, ..., r,\ l = 0, 1, ..., (t-1),\ g$ is a monotonically increasing function, y_{ijl} is the gray-intensity level at the l^{th} dilution of the j^{th} replicate for the i^{th} sample, c_i is the protein concentration level of the i^{th} sample and ϵ_{ijl} is an error term. The function g may be assumed to have either a parametric form or a nonparametric form. The typical parametric model, called the Sigmoidal model, takes the form:

$$y_{ijl} = \beta_1 + \frac{\beta_2}{1 + e^{-\beta_3(x_i - l)}} + \epsilon_{ijl}, \tag{1.1}$$

where $\beta_1, \beta_2, \beta_3 > 0,\ y_{ijl}$ is the gray-intensity level, x_i is the logarithm of the protein concentration level with base 2 $\left(x_i = log_2(c^i)\ \text{ or }\ x_i - l = log_2(c^i/2^l)\right)$ and ϵ_{ijl} is an error term with $E(\epsilon_{ijl}) = 0$ and $var(\epsilon_{ijl}) = \sigma^2$. We notice

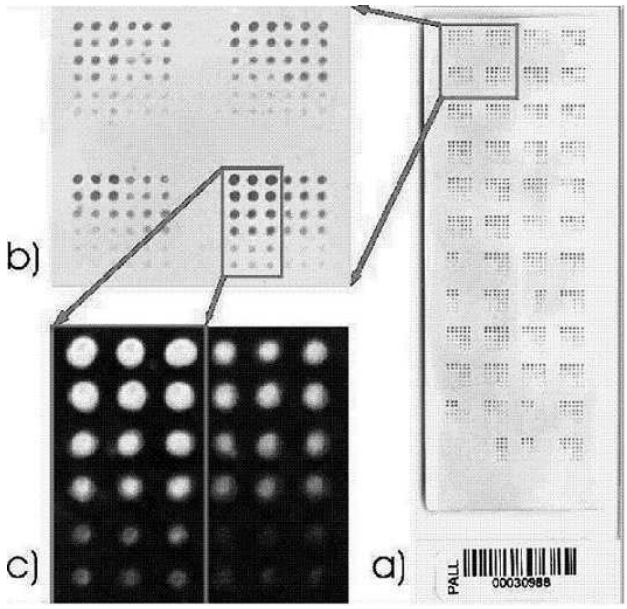

Figure 1.1: Protein lysate array design. Source: [10]. a) Layout of the protein lysate array with $s = 96$ biological samples, $r = 3$ replicates, and $t = 6$ dilution levels. b) Magnification corresponding to eight samples. c) Magnification corresponding to two samples.

that $\beta_1 = \lim_{x_i \to -\infty} E(y_{ijl})$ and $\beta_1 + \beta_2 = \lim_{x_i \to \infty} E(y_{ijl})$. Therefore, we have rough interpretations for the parameters: β_1 is the lowest intensity level without noise, and β_2 is the increment from the lowest intensity level to the highest intensity level (representing the saturation of intensity). This model looks like a general nonlinear regression model where y_{ijl} is a dependent variable and x_i is an explanatory variable. However, unlike a usual regression model, we do not observe x_i, but we need to estimate it. Under the model (1.1), our goal is to estimate the common curve parameters, β_1, β_2, and β_3, and more importantly, the logarithm of the protein concentration, x_i, even though our final interest is the difference in the logarithms of the protein concentrations, $x_i - x_i^*$, that is equivalent to $log_2(c_i/c_i^*)$ where $i, i^* = 1, 2, ..., s,\ i \neq i^*$.

In reality, the number of replicates, r, and the number of dilution levels, t, are limited by time and cost constraints as compared with the number of biological samples, s, particularly when we are interested in measuring the protein concentration levels of a large number of biological samples. A sensible large sample framework is to consider asymptotics as both n $(= rt)$ and s increase, which leads to a problem where the number of unknown parameters grows with the available data points. In this case, the

parameter estimators, such as the maximum likelihood estimator, require a non-standard asymptotic analysis ([11]).

Most prior work on protein array quantification is based on the assumption that the errors are independent. However, when we examine the residuals obtained under the *i.i.d.* error assumption, this assumption appears very questionable. The nature of the experiment warrants the existence of correlation, too. The repeated measurements of each biological sample are likely to be correlated. Also the measurements within dilution series come from the same replicate of the same biological sample, and are thereby likely to have correlation.

1.3 Prior Work on Protein Lysate Array Quantification

The protein lysate array is a relatively new technology for measuring the protein concentration ratios in a large number of biological samples, and a limited amount of statistical literature on this topic is available.

One approach for measuring the protein concentration ratios is introduced in [12], where separate estimation for each biological sample is employed. They propose two methods for estimating the protein concentration levels from 1440-spot lysate arrays with 80 samples, three replicates, and six 2-fold dilutions. In the first method, the median value of the the gray-intensity levels for the three replicates is obtained at each dilution for each biological sample. Suppose that the median observation at the l^{th} dilution for the i^{th} sample is denoted by $y_{i.l}$, typically on the log scale of the intensity measurements. They assume the following linear model for each sample,

$$y_{i.l} = \gamma_1 + \gamma_2\, l + \epsilon_{i.l},$$

where $l = 1, 2, ..., 6$, and apply the least squares regression to estimate γ_1 and γ_2. To compensate for the lack of robustness in the first method, they propose the second method that uses all 18 measurements (3 replicates $\times$ 6 dilution levels) and fits a linear model using a robust estimation for each sample. For the i^{th} sample, they assume the following linear model,

$$y_{ijl} = \gamma_1 + \gamma_2\, l + \epsilon_{ijl},$$

where $j = 1, 2, 3,\ l = 1, 2, ..., 6$, and then apply a robust method to estimate γ_1 and γ_2. For both methods, the logarithm of the protein concentration ratio is estimated as the distance between two fitted lines if the two lines

are parallel. If the two lines are not parallel, the log-ratio is estimated as the weighted distance between two fitted lines at each dilution. The weight depends on the dispersion at that particular dilution: a dilution that has a higher dispersion of three replicates is assigned a smaller weight and a dilution that has a lower dispersion of three replicates is assigned a larger weight.

However, it is generally believed that the intensity has a nonlinear relationship with the logarithm of true protein concentration level. Specifically, the S-shape is expected because of both the background noise at the low protein concentration level and the saturation intensity at the high protein concentration level. In addition, it would be more reasonable to assume a common curve for all the biological samples, because the identical antibody to recognize the target protein is used for all the samples, and thus the target protein in different samples reacts similarly to the antibody. In this case, a joint estimation of the common curve using all the measurements of all the samples will be more efficient than the separate estimation of the curve for each sample.

A joint estimation assuming a common parametric curve is discussed in [10]. They postulate several polynomial models and the Sigmoidal model for dependence between the intensity level and the protein concentration level. Among these models, they acknowledge that the Sigmoidal model generally yields the best fit to lysate data. We introduce the procedure with the Sigmoidal model, but the procedure with the polynomial model is similar. The Sigmoidal model introduced in the previous section takes the form:

$$y_{ijl} = \beta_1 + \frac{\beta_2}{1 + e^{-\beta_3(x_i - l)}} + \epsilon_{ijl},$$

where y_{ijl} is the gray-intensity level and x_i is the logarithm of the protein concentration. Their proposed method is based on the least squares estimation: after the initial estimates of the concentration levels are obtained (typically assuming a linear model), the estimations of the curve parameters $\underline{\beta}$, and of the concentration levels x_i, are iteratively performed via nonlinear least squares. The details are given below.

1. Consider a simple linear model, $y_{ijl} = \beta_0(x_i - l) + \epsilon_{ijl}$, where $\beta_0 > 0$, which can be rearranged as follows:

$$\begin{aligned} y_{ijl} &= \beta_0(x_i - l) + \epsilon_{ijl} \\ &= \gamma_i - \beta_0 l + \epsilon_{ijl}, \end{aligned}$$

with $\gamma_i = \beta_0 x_i$. Then we can find the least squares estimators, $\hat{\beta}_0$ and $\hat{\gamma}_i,\ i = 1, 2, ..., s$, that minimize

$$\sum_{i=1}^{s}\sum_{j=1}^{r}\sum_{l=0}^{t-1}(y_{ijl} - \gamma_i + \beta_0 l)^2,$$

from which $\hat{x}_i$ is obtained as $\hat{\gamma}_i/\hat{\beta}_0,\ i = 1, 2, ..., s$.

2. Given x_i as $\hat{x}_i$, find the least squares estimators, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ that minimize

$$\sum_{i=1}^{s}\sum_{j=1}^{r}\sum_{l=0}^{t-1} w_i^2\big(y_{ijl} - \beta_1 - \frac{\beta_2}{1 + e^{-\beta_3(\hat{x}_i - l)}}\big)^2,$$

where w_i can be set to zero to exclude the unreliable data points at the spot quantification stage. Given β_1, β_2, and β_3 as $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$, respectively, find the least squares estimator, $\hat{x}_i$ that minimizes

$$\sum_{j=1}^{r}\sum_{l=0}^{t-1} w_i^2\big(y_{ijl} - \hat{\beta}_1 - \frac{\hat{\beta}_2}{1 + e^{-\hat{\beta}_3(x_i - l)}}\big)^2.$$

3. Iterate step 2 until the change in $\sum_{i=1}^{s}\sum_{j=1}^{r}\sum_{l=0}^{t-1} w_i^2(y_{ijl} - \hat{\beta}_1 - \frac{\hat{\beta}_2}{1+e^{-\hat{\beta}_3(\hat{x}_i - l)}})^2$ after an iteration becomes small enough.

In this method, however, the asymptotic behaviors of the final estimates are not well studied, and the optimization over $3 + s$ parameters is subject to the risk of missing the global minimum. Besides that, the model does not take into consideration the dependence among the measurements, which seems quite necessary when analyzing the lysate array data according to our examination of real data.

Another approach, a nonparametric approach, is discussed in [13]. Without specifying a parametric form they propose the model,

$$y_{ijl} = g(x_i - l) + \epsilon_{ijl},$$

where g is a monotonically increasing function and the median of ϵ_{ijl} is assumed to be zero. After the initial estimates of the concentration levels are obtained assuming a linear model, the estimations of the link function, g, and of the concentration levels x_i, are iteratively performed. The details are given below.

1. Consider a linear model, $y_{ijl} = \alpha_0 + \alpha_1(x_i - l) + \epsilon_{ijl}$. The estimator

$\hat{\alpha}_0$ is obtained as the minimum of all the intensity measurements, $\hat{\alpha}_1$ is the median slope over all the dilution series, and $\hat{x}_i$ is the median of $\{(y_{ijl} - \hat{\alpha}_0 + \hat{\alpha}_1 l)/\hat{\alpha}_1 : \ j = 1, 2, ..., r, \ l = 0, 1, ..., (t-1)\}$.

2. Given x_i as $\hat{x}_i$, find $\hat{g}$ that, with a monotonicity constraint on g, minimizes
$$\sum_{i=1}^{s}\sum_{j=1}^{r}\sum_{l=0}^{t-1} |y_{ijl} - g(\hat{x}_i - l)| + \lambda \, max_x |g^{''}(x)|,$$
where λ is a smoothing parameter. Now given g as $\hat{g}$, find $\hat{x}_i$ that minimizes
$$\sum_{j=1}^{r}\sum_{l=0}^{t-1} |y_{ijl} - \hat{g}(x_i - l)|.$$

3. Iterate step 2 until the change in $\sum_{i=1}^{s}\sum_{j=1}^{r}\sum_{l=0}^{t-1} |y_{ijl} - \hat{g}(\hat{x}_i - l)|$ after an iteration becomes small enough.

Even though the nonparametric approach of [13] is flexible without assuming any parametric form, this method also involves estimating the parameters of increasing dimensions because x_i and g are estimated simultaneously using all the measurements, which makes asymptotic inference of the final estimators difficult. In addition, the dependence structure of data has not been taken into account.

[14] discuss the asymptotic behavior of M-estimators when the dimension of the parameter space increases with the sample size. They show that under certain regulatory conditions, M-estimators are consistent and approximately normal if the dimension of the parameter space grows at a controlled rate relative to the sample size. However, we find that it is hard to apply their results directly to the protein lysate array with the Sigmoidal model due to the difference in the convergency rates between the estimates of x_i and $\beta_k, \ k = 1, 2, 3$. Note that we expect $|\hat{x}_i - x_i| = O_p(\frac{1}{\sqrt{rt}})$ and $|\hat{\beta_k} - \beta_k| = O_p(\frac{1}{\sqrt{srt}})$, because only rt measurements are used to estimate x_i, while all srt measurements are used to estimate $\beta_k, \ k = 1, 2, 3$. Then the matrix that divides the elements of the Hessian matrix by the sample size, srt, is not invertible and one of the regulatory conditions is not satisfied, preventing us from directly applying their results to the protein lysate array.

A protein lysate array measures the relative concentration levels of a particular protein in many samples. Hence, in order to measure the levels of more than one protein, people use a set of identically spotted arrays. Most studies focus on the estimation of the concentration levels of a particular protein by using within array information. Recent work of [15] proposes

a model that takes into consideration the sample effect and the correlation between proteins from several arrays. Additional studies are reported in [2], [4] and [16]. However, neither asymptotic behaviors nor specific dependence structures among the measurements within the same sample have not been studied in those papers.

The asymptotic analysis plays an important role in statistics. It provides a good basis for understanding the behaviors of an estimator and helps make large sample inferences for statistical models. As far as we know, there is not yet any asymptotic theory on protein quantification methods that can be used to justify approximate inference procedures statistically. Chapter 2 aims to propose a method that guarantees a consistent estimator of the protein concentration in the protein lysate array where the dimension of the parameters increases with the sample size.

Most prior work on protein array quantification is based on the $i.i.d.$ error assumption, but the real data analysis often exhibits evidence against this assumption. Misspecified models may lead to incorrect inference on an estimator. In Chapter 3, we introduce a model that allows for the dependence structure, and propose a method that can incorporate the complexity of the correlation structure of data.

CHAPTER 2

MULTI-STEP PROTEIN LYSATE ARRAY QUANTIFICATION METHOD

2.1 Introduction

The protein lysate array is a developing technology for estimating the protein concentration ratios in a large number of biological samples. Each sample is serially diluted by a certain factor, spotted on a nitrocellulose-coated slide in multiple replicates, and then bound with an antibody to measure the amount of the protein of interest. As a result, a gray-level image is obtained at each dilution level of each replicate for each sample. The dilution, as well as the replicates, give several measurements for each sample, which is a key characteristic of the protein lysate array from the quantification point of view. In reality, the number of replicates, r, and the number of dilution levels, t, are limited by time and cost constraints as compared with the number of biological samples, s, particularly when we are interested in measuring the protein concentration levels of a large number of biological samples. A sensible large sample framework is to consider asymptotics as both n $(= rt)$ and s increase, which leads to a problem where the number of unknown parameters grows with the available data points. It makes the analysis of the protein lysate array a problem on parameters of increasing dimensions. When we have this problem, finding a consistent estimator is not straightforward. This statistical issue has not been dealt with in previous work on protein lysate arrays, and it motivates our work. We propose a multi-step least squares procedure as a modification of earlier methods of protein quantification. The multi-step procedure applies the least squares estimation to biological samples in small groups, and then uses a pooled curve parameter estimate to recover efficiency.

The rest of this chapter is organized as follows. First, in Section 2.2, we propose our multi-step procedure with two subtypes, depending on the methods of pooling. In Section 2.3, we show the consistency and the asymptotic normality of the protein concentration estimates. In Section 2.4, we evaluate the finite sample performance of the proposed procedure relative

to the existing method of least squares, based on simulation studies. We provide a confidence interval for the relative concentration level in the same section. Real data analysis based on two different lysate arrays is given in Section 2.5. Section 2.6 concludes the chapter.

2.2 Multi-Step Procedure

Throughout this chapter we assume the Sigmoidal model (1.1) introduced in Section 1.2 for the relationship between the intensity level and the protein concentration, which seems reasonable because the Sigmoidal model reflects key characteristics of the lysate array data with background noise at the lower end and the saturation at the higher end. Consider model (1.1) again:

$$y_{ijl} = \beta_1 + \frac{\beta_2}{1 + e^{-\beta_3(x_i - l)}} + \epsilon_{ijl},$$

where $\beta_1, \beta_2, \beta_3 > 0$. As we previously mentioned, β_1 and β_2 are interpreted as the lowest intensity level and the increment from the lowest intensity level to the saturation, respectively. We denote the curve parameters by $\underline{\beta} = (\beta_1, \beta_2, \beta_3)'$. A suspicion that β_1 and β_2 could be inversely correlated leads us to consider a reparameterization with a new parameter, $\gamma = (\beta_1 + \beta_2)$, and a new model,

$$y_{ijl} = \beta_1 + \frac{\gamma - \beta_1}{1 + e^{-\beta_3(x_i - l)}} + \epsilon_{ijl}, \tag{2.1}$$

where $\gamma > \beta_1 > 0$ and $\beta_3 > 0$.

Based on model (1.1) and model (2.1), we now propose a multi-step procedure that is simpler both in theory and in computations, while achieving full asymptotic efficiency of the concentration level estimates.

2.2.1 Procedure Details

Our modification has two components: a divide and conquer component and a pooling component. In particular, we choose a small value of k, divide the s biological samples into s/k (or its integer part) groups, use the least squares method to estimate the parameters in each group, and then pool the curve parameters from all groups, and finally, estimate x_i for each sample based on the pooled estimate of the curve parameters. The details of the modified procedure are given as follows.

Step 1 Grouping:

We combine k samples into one group, resulting in s/k groups and krt measurements per group. In particular, we first find the median of rt measurements for each biological sample and divide the s samples into k categories so that the first category consists of the s/k samples that have the smallest medians, the second category consists of the s/k samples that have the next smallest medians, and so on. The last category consists of the s/k samples that have the largest medians. Then, randomly select one sample from each category to form a group, where each group is comprised of k samples.

Step 2 Estimating parameters for each group:

Denote the curve parameters and the concentration levels of the q^{th} group by $\underline{\beta}^q = (\beta_1^q, \beta_2^q, \beta_3^q)'$ and $\underline{x}^q = (x_1^q, ..., x_k^q)'$, respectively, where $q = 1, \cdots, s/k$. In addition, we use the notation y_{ijl}^q for the measurements of the i^{th} sample in the q^{th} group. Then, find the least squares estimates, $\hat{\underline{\beta}}^q$ and $\hat{\underline{x}}^q$, that minimize

$$\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=0}^{t-1}\big(y_{ijl}^q - \beta_1^q - \frac{\beta_2^q}{1+e^{-\beta_3^q(x_i^q-l)}}\big)^2. \tag{2.2}$$

Step 3 Pooling the curve parameter estimates:

Combine $\hat{\underline{\beta}}^q$ linearly with an appropriate weight matrix so that the pooled estimate is

$$\hat{\underline{\beta}}^{(c)} = (\hat{\beta}_1^{(c)}, \hat{\beta}_2^{(c)}, \hat{\beta}_3^{(c)})' = \sum_q V^q \hat{\underline{\beta}}^q,$$

where two types of weight matrices, V^q, will be detailed in Section 2.2.2.

Step 4 Estimating the protein concentration level:

Given $\hat{\underline{\beta}}^{(c)}$, find the concentration estimate, $\tilde{x}_i$, for the i^{th} sample by minimizing

$$\sum_{j=1}^{r}\sum_{l=0}^{t-1}\big(y_{ijl} - \hat{\beta}_1^{(c)} - \frac{\hat{\beta}_2^{(c)}}{1+e^{-\hat{\beta}_3^{(c)}(x_i-l)}}\big)^2. \tag{2.3}$$

Because our proposed method employs an individual estimation for each group, it might seem to use limited information from data. Pooling the

estimates in Step 3, however, enables us to bring back information across the groups. There is a trade-off between Step 2 and Step 3 as the size of k varies. If k is small, the within-group variability is high, and thus the curve parameter may not be estimated well in Step 2, but more groups are available for pooling in Step 3 to regain efficiency. Based on our empirical experience, $k = 3$ or 4 works reasonably well. In theory, any finite value of k leads to the same asymptotic efficiency, but computational satiability is easier to attain at a small value of k.

The strategy that we use for grouping in Step 1 allows us to estimate the group-based curve parameter more stably. We divide the biological sample into k categories according to the median of the measurements and then randomly select one sample from each category to make a group. Then each group can have measurements that cover a wider range of data, which can make the group-based estimate more stable.

Our proposed method involves two optimization problems: one with $3+k$ parameters in Step 2 and the other one with one parameter (at each time) in Step 4, whereas the original least squares method solves an optimization problem with $3 + s$ parameters. The lower dimensional parameter space in our proposed method reduces the risk of missing the global minimum.

So far we illustrated our method using the original model (1.1). With regard to the reparameterized model (2.1), every step will be identical except $\beta_2^q = \gamma^q - \beta_1^q$.

2.2.2 Weight Matrices

A proper weight matrix is needed to combine $\underline{\hat{\beta}}^q$, $q = 1, 2, ..., s/k$, and two weight matrices are discussed in this section. We only use the original Sigmoidal model in this section, but we can easily apply the results to the reparameterized Sigmoidal model.

The first weight matrix, $V^q, q = 1, ..., s/k$, minimizes the trace of $var\big(\underline{\hat{\beta}}^{(c)}\big)$ subject to $\sum_q V^q = I_3$, where V^q is a 3×3 matrix and I_3 is the 3×3 identity matrix. By using a result of [17], we obtain V^q

$$V^q = (\sum_{q=1}^{s/k} \Omega_q^{-1})^{-1}\ \Omega_q^{-1},$$

where Ω_q is the variance-covariance matrix of $\underline{\hat{\beta}}^q$. Consider $\underline{\hat{\beta}}^*$ that mini-

mizes the following function $N(\underline{\beta})$,

$$\arg\min_{\underline{\beta}} N(\underline{\beta}) = \arg\min_{\underline{\beta}} \sum_{q=1}^{s/k} (\underline{\beta} - \underline{\hat{\beta}}^q)' \ \Omega_q^{-1} \ (\underline{\beta} - \underline{\hat{\beta}}^q).$$

It can be easily shown that

$$\underline{\hat{\beta}}^* = \sum_{q=1}^{s/k} (\sum_{q=1}^{s/k} \Omega_q^{-1})^{-1} \ \Omega_q^{-1} \underline{\hat{\beta}}^q.$$

Therefore, minimizing $N(\underline{\beta})$ leads to the exact same weight matrix as does the trace-minimizing criterion.

Since V^q is not necessarily a diagonal matrix, the variances of the curve parameter estimates are treated jointly. However, the curve parameters, β_1, β_2, and β_3, are usually in different scales and $var(\hat{\beta}_3)$ is expected to be smaller than $var(\hat{\beta}_1)$ and $var(\hat{\beta}_2)$. To treat each variance separately, we may consider a diagonal weight matrix,

$$V^q = \begin{pmatrix} C_1 & 0 & 0 \\ 0 & C_2 & 0 \\ 0 & 0 & C_3 \end{pmatrix},$$

where $C_m = \left(\sum_{q=1}^{s/k} \left(var(\hat{\beta}_m^q) \right)^{-1} \right)^{-1} \left(var(\hat{\beta}_m^q) \right)^{-1}$, $m = 1, 2, 3$. We remark that $var(\hat{\beta}_m^q)$ is the diagonal element of Ω_q, which implies that the component-wise minimization treats each variance component separately. In case an estimated covariance matrix, $\hat{\Omega}_q$, is singular, which poses problems in the computation of the weight matrix, we use a weight of zero for the corresponding $\underline{\hat{\beta}}^q$.

2.2.3 Variance-Covariance Matrix of Curve Parameter Estimates

In this section, we examine the covariance structure of estimated curve parameters. The variance-covariance matrix in the original Sigmoidal model, $cov(\underline{\hat{\beta}}^q)$, where $\underline{\hat{\beta}}^q = (\ \hat{\beta}_1^q, \ \hat{\beta}_2^q, \ \hat{\beta}_3^q \ )'$, is considered first and then the variance-covariance matrix in the reparameterized model, $cov(\underline{\hat{\Gamma}}^q)$, where $\underline{\hat{\Gamma}}^q = (\ \hat{\beta}_1^q, \ \hat{\gamma}^q, \ \hat{\beta}_3^q \ )'$, is derived by modifying $cov(\underline{\hat{\beta}}^q)$.

Consider the original Sigmoidal model with the objective function (2.2) and denote the vector of the least squares estimators by $\underline{\hat{\theta}}^q = (\ \hat{\beta}_1^q, \ \hat{\beta}_2^q, \ \hat{\beta}_3^q, \ \hat{x}_1^q, \ \ldots, \hat{x}_k^q \ )'$.

Then $\hat{\underline{\theta}}^q$ is the solution to

$$\sum_{j=1}^{r}\sum_{l=0}^{t-1}\psi(y^q_{1jl},\ y^q_{2jl},\ \ldots,y^q_{kjl},\ \underline{\theta}^q) = \ \underline{0}\,, \tag{2.4}$$

where ψ is vector-valued with the $(3+k)$ components:

$$[\psi]_1 = -2\ \sum_{i=1}^{k} e^q_{ijl},$$

$$[\psi]_2 = -2\ \sum_{i=1}^{k} e^q_{ijl}\ \frac{1}{1+e^{-\beta^q_3(x^q_i-l)}}\ ,$$

$$[\psi]_3 = -2\ \sum_{i=1}^{k} e^q_{ijl}\ \frac{\beta^q_2(x^q_i-l)e^{-\beta^q_3(x^q_i-l)}}{\left(1+e^{-\beta^q_3(x^q_i-l)}\right)^2},$$

$$[\psi]_{(u+3)} = -2\ e^q_{ijl}\ \frac{\beta^q_2\beta^q_3 e^{-\beta^q_3(x^q_u-l)}}{\left(1+e^{-\beta^q_3(x^q_u-l)}\right)^2}\ ,\ u=1,2,...,k,$$

where $e^q_{ijl} = y^q_{ijl} - \beta^q_1 - \frac{\beta^q_2}{1+e^{-\beta^q_3(x^q_i-l)}}$. By the sandwich formula for the standard generalized estimating equations (GEE, [18]), we have

$$v_q = var\left(\hat{\underline{\theta}}^q\right) = B_q^{-1}M_qB_q^{-1},$$

where B_q and M_q are symmetric matrices of the following forms:

$$B_q = E\sum_{j=1}^{r}\sum_{l=0}^{t-1}\frac{\partial\psi}{\partial\underline{\theta}^q} = 2r\ \left[B_{vw}\right]_{1\,\le\,v,\,w\,\le\,(3+k)} \quad \text{and}$$

$$M_q = var\sum_{j=1}^{r}\sum_{l=0}^{t-1}\psi = 4r\sigma^2\ \left[M_{vw}\right]_{1\,\le\,v,\,w\,\le\,(3+k)}.$$

In Appendix A, we give the explicit expressions of B_{vw} and M_{vw}. Once v_q is calculated, the variance-covariance matrix of $\hat{\underline{\beta}}^q$ can be obtained as

$$\Omega_q = var(\hat{\underline{\beta}}^q) = Tv_qT',$$

where T is the following $3\times(k+3)$ matrix:

$$T = \begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \end{pmatrix}. \tag{2.5}$$

In practice, the true values of $\underline{\beta}^q$, $\underline{x}^q$, and σ^2 are unknown and have to be replaced by their estimated values. We use $\hat{\underline{\beta}}^q$ and $\hat{\underline{x}}^q$, the estimates from Step 2, to estimate σ^2 by

$$\widehat{(\sigma^2)}^q = \sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=0}^{t-1}\big(y_{ijl}^q - \hat{\beta}_1^q - \frac{\hat{\beta}_2^q}{1+e^{-\hat{\beta}_3^q(\hat{x}_i^q - l)}}\big)^2 \Big/ \Big(krt-(3+k)\Big).$$

Now consider $cov(\hat{\underline{\Gamma}}^q) = cov\big((\hat{\beta}_1^q, \hat{\gamma}^q, \hat{\beta}_3^q)'\big)$ in the reparameterized Sigmoidal model. Since $\gamma = (\beta_1 + \beta_2)$, we have

$$cov\big(\hat{\underline{\Gamma}}^q\big) = cov\big((\hat{\beta}_1^q, \hat{\gamma}^q, \hat{\beta}_3^q)'\big) = cov\big(Z\ (\hat{\beta}_1^q, \hat{\beta}_2^q, \hat{\beta}_3^q)'\big) = cov\big(Z\ \hat{\underline{\beta}}^q\big) = Z\ cov\big(\hat{\underline{\beta}}^q\big)\ Z',\ \text{and}$$

$$\widehat{cov\big(\hat{\underline{\Gamma}}^q\big)} = Z\ \widehat{cov\big(\hat{\underline{\beta}}^q\big)}\ Z',$$

where

$$Z = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

2.3 Asymptotic Properties

We first show the consistency of $\hat{\underline{\beta}}^{(c)}$, and then establish the asymptotic normality of $\tilde{x}_i$. We assume that the errors are independent and identically distributed in Model (1.1). Only the original Sigmoidal model is considered in this section, but the results can easily be modified for the reparameterized model.

Consider the objective function (2.2) for the q^{th} group in Step 2. We continue to use the notation $\hat{\underline{\theta}}^q = \big((\hat{\underline{\beta}}^q)', (\hat{\underline{x}}^q)'\big)'$ for the least squares estimates that solve (2.4), and denote the true parameters by $\underline{\theta}_o^q = \big((\underline{\beta}_o)', (\underline{x}_o^q)'\big)' = (\beta_{1o},\ \beta_{2o},\ \beta_{3o},\ x_{1o}^q, ...,\ x_{ko}^q)'$. Following [11], the Bahadur representation for $\hat{\underline{\theta}}^q$ is given by

$$\sqrt{n}\ (\hat{\underline{\theta}}^q - \underline{\theta}_o^q) = C^{-1}\ D + \sqrt{n}\ \underline{R}_n^q\ ,$$

where

$$n = rt,$$

$$\underline{R}_n^q = o_p(n^{-1/2}),$$

$$C = -\frac{1}{n}\sum_{j=1}^{r}\sum_{l=0}^{t-1}\dot{\psi}(y^q_{1jl},...,y^q_{kjl},\underline{\theta}^q_o),$$

$$D = \frac{1}{\sqrt{n}}\sum_{j=1}^{r}\sum_{l=0}^{t-1}\psi(y^q_{1jl},...,y^q_{kjl},\underline{\theta}^q_o),$$

and $\dot{\psi}(y^q_{1jl},...,y^q_{kjl},\underline{\theta}^q_o) = \big[\partial\psi(y^q_{1jl},...,y^q_{kjl},\underline{\theta}^q)/\partial(\underline{\theta}^q)'\big]|_{\underline{\theta}^q=\underline{\theta}^q_o}$. Then it follows that

$$\underline{\hat{\beta}}^q = \underline{\beta}_o + \frac{1}{n}\,T\,\sum_{j=1}^{r}\sum_{l=0}^{t-1}\varphi_{IF}(y^q_{1jl},...,y^q_{kjl},\underline{\theta}^q_o) + T\,\underline{R}^q_n,$$

where T is a $3\times(3+k)$ matrix given in (2.5), and $\varphi_{IF}(y^q_{1jl},...,y^q_{kjl},\underline{\theta}^q_o) = C^{-1}\ \psi(y^q_{1jl},...,y^q_{kjl},\underline{\theta}^q_o)$. After combining the group-specific estimates, we have

$$\begin{aligned}\underline{\hat{\beta}}^{(c)} &= \sum_{q=1}^{s/k} V^q\underline{\hat{\beta}}^q \\ &= \underline{\beta}_o + \frac{1}{n}\sum_{q=1}^{s/k}\sum_{j=1}^{r}\sum_{l=0}^{t-1}V^q\,T\,\varphi_{IF}(y^q_{1jl},...,y^q_{kjl},\ \underline{\theta}^q_o) + \sum_{q=1}^{s/k}V^q\,T\,\underline{R}^q_n.\end{aligned}$$

Under conditions (A1)–(A3) as shown in Appendix B, we have the following lemma.

Lemma 1. If conditions (A1)–(A3) are satisfied and $n = rt$, then

$$||\underline{\hat{\beta}}^{(c)} - \underline{\beta}_o|| = O_p\big((ns)^{-1/2}\big)\ \text{ as }\ n\to\infty\ \text{ and }\ s\to\infty.$$

Lemma 1 holds for both trace minimization criterion and component-wise minimization criterion. The proof is also given in Appendix B.

Now consider the objective function (2.3) for the i^{th} sample in Step 4. The least squares estimate, $\tilde{x}_i$, solves

$$\sum_{j=1}^{r}\sum_{l=0}^{t-1}\phi(y_{ijl},\ \underline{\hat{\beta}}^{(c)},\ x_i - l) = 0,$$

where

$$\phi(y_{ijl},\ \underline{\hat{\beta}}^{(c)}, x_i - l) = -2\Big(y_{ijl} - \hat{\beta}^{(c)}_1 - \frac{\hat{\beta}^{(c)}_2}{1+e^{-\hat{\beta}^{(c)}_3(x_i-l)}}\Big)\frac{\hat{\beta}^{(c)}_2\hat{\beta}^{(c)}_3 e^{-\hat{\beta}^{(c)}_3(x_i-l)}}{\big(1+e^{-\hat{\beta}^{(c)}_3(x_i-l)}\big)^2}.$$

Then we have the following Bahadur representation for $\tilde{x}_i$:

$$\sqrt{n}\,(\tilde{x}_i - x_{io}) = \frac{1}{\sqrt{n}} \sum_{j=1}^{r} \sum_{l=0}^{t-1} \varphi_{IC}(y_{ijl},\, \underline{\hat{\beta}}^{(c)},\, x_{io} - l) + o_p(1),$$

where

$$\varphi_{IC}(y_{ijl},\, \underline{\hat{\beta}}^{(c)}, x_{io} - l) = \Big(- \frac{1}{n} \sum_{j=1}^{r} \sum_{l=0}^{t-1} \dot{\phi}(y_{ijl},\, \underline{\hat{\beta}}^{(c)}, x_{io} - l) \Big)^{-1} \phi(y_{ijl},\, \underline{\hat{\beta}}^{(c)},\, x_{io} - l),$$

and

$$\dot{\phi}(y_{ijl},\, \underline{\hat{\beta}}^{(c)}, x_{io} - l) = \big[d\phi(y_{ijl},\, \underline{\hat{\beta}}^{(c)}, x_i - l)/dx_i \big]|_{x_i = x_{io}}.$$

Theorem 1. If conditions (B1) given in Appendix C is satisfied, then

$$\sqrt{n}(\tilde{x}_i - x_{io})/s_n \xrightarrow{d} N(0,1) \;\text{ as }\; n \to \infty \text{ and } s \to \infty,$$

where $s_n^2 = t\sigma^2 \left[\sum_{l=0}^{t-1} \frac{(\beta_{2o})^2 (\beta_{3o})^2 e^{-2\beta_{3o}(x_{io}-l)}}{\left(1+e^{-\beta_{3o}(x_{io}-l)}\right)^4} \right]^{-1}$. The proof is given in Appendix C. We note that the asymptotic variance of $\tilde{x}_i$ is the same as that under a model with known curve parameters.

Because $\underline{\hat{\beta}}^{(c)}$ converges faster than the rate $n^{-1/2}$, the estimates of x_i in Step 4 are as efficient as in a problem with known curve parameters. For this reason, the estimator, $\tilde{x}_i$ achieves its full asymptotic efficiency. We expect the same to be true for the least squares estimator of [10], but the result is harder to verify. The single-stage method in [10] involves estimating the parameters of increasing dimensions because $x_i, i = 1, 2, ..., s$, and $\beta_m, m = 1, 2, 3$, are estimated simultaneously using all the measurements, which makes asymptotic inference of the final estimators difficult.

2.4 Simulation Studies

In this section, we carry out several simulation studies to evaluate the finite sample performance of the proposed multi-step procedure in comparison with a single-step least squares procedure on all the samples used in [10] (to be denoted M-S hereafter). Our proposed procedure can be divided into two subtypes depending on the methods of pooling. The first subtype, M-T, uses the weight matrix that minimizes the trace of the variance-covariance matrix, and the second subtype, M-C, uses the weight matrix that minimizes the diagonal elements of the variance-covariance matrix. We have not found

a big difference in the estimation results between the original Sigmoidal model and the reparameterized Sigmoidal model, so we present only the result based on the original model. In addition, we provide a confidence interval for the relative concentration level by using the bootstrap methods and the asymptotic theory developed in Section 2.3.

2.4.1 Algorithm Details

The proposed multi-step method combines k biological samples into one group. In the simulation studies, we use three different k values: $k = 2, 3$, or 4. When the estimated variance-covariance matrix is singular in Step 3, which makes the computation of the weight matrix impossible, we use a weight of zero for the corresponding $\underline{\hat{\beta}}^q$. We use the R function *solve* with the default tolerance level to determine singularity. Steps 2 and 4 of our proposed method require nonlinear optimization. The R function *optim* is used for the implementation with the following starting values.

In Step2:

1. As the starting value of β_1, use the minimum of the intensity measurements.

2. As the starting value of β_2, use the range of the intensity measurements. (as the starting value of γ in the reparameterized model, use the maximum of the measurements).

3. As the starting value of β_3, simply use a small ad hoc value 0.01.

4. As the starting values of x_i, use the initial estimates of the concentration levels proposed in [10]. We center those values to have median zero.

In Step4: As the starting values of x_i, use $\hat{x}_i$, the estimates from Step 2.

Note that the Sigmoidal model has lower and upper asymptotes, which means that if the x_i's lie in the tail area, they can hardly be distinguished from the intensity measurements. For this reason, we Winsorize the final estimates $\tilde{x}_i$ in both tails. Specifically, we compute

$$f(x) = \hat{\beta}_1^{(c)} + \frac{\hat{\beta}_2^{(c)}}{1 + e^{-\hat{\beta}_3^{(c)} x}},$$

and then find $x^* > 0$ such that $f'(x^*) = z$, for a slope threshold value z. Then all the points $\tilde{x}_i$ that satisfy $\tilde{x}_i > x^*$ are Winsorized to x^*, and all the points $\tilde{x}_i$ that satisfy $\tilde{x}_i < -x^*$ are Winsorized to $-x^*$. The slope threshold, z, is chosen as 0.05% of the range of the intensity measurements. This choice of z is ad hoc, supported by our empirical experience with lysate array data.

For the single-stage method used in [10], the same starting values of $\beta_1, \beta_2, \beta_3$ and x_i as given above are used and the R function *optim* is employed in the implementation. Also the same Winsorization rule is applied.

2.4.2 Assessment Criteria

The aim of the simulation studies is to evaluate how well the relative concentration levels are estimated. [13] provide a relevant summary and we use the same idea with suitable modifications. We perform M simulation trials, and for each trial we find as a reference the biological sample whose true concentration level is the median of s true levels. Then, for that trial compute D_{tot} as defined by

$$D_{tot} = \sum_{i=1}^{s} |\tilde{D}_i - D_i|,$$

where $\tilde{D}_i = \tilde{x}_i - \tilde{x}_{ref}$, $D_i = x_i - x_{ref}$, $\tilde{x}_i$ and $\tilde{x}_{ref}$ are the estimates for the concentration levels of the i^{th} sample and of the reference sample, respectively, and x_i and x_{ref} are their true values. The smaller D_{tot} is, the more desirable the result is.

2.4.3 Simulation Case 1

In the first case, we intend to mimic the result of the real data analysis performed in [10] in order to reflect a realistic situation. The lysate array for the protein $pThr308AKT$ has been analyzed in [10], and we use the result as a reference when generating simulation data sets. The array $pThr308AKT$ is accessible online (http://www.cs.tut.fi/∼tabus/lysate/).

We generate data from Model (1.1) with $(\beta_1, \beta_2, \beta_3) = (10, 6, 0.5)$. A total of 96 samples are used with $r = 3$ dilution series of length $l = 8$ for each sample. The log scale of the true concentration levels, x_i, are generated from the uniform distribution on the interval $(-1.5,\ 8.5)$, and then all the dilution series, $(x_i - l),\ l = 0, 1, ..., 7$, range within the interval $(-8.5,\ 8.5)$. The independent error terms, ϵ_{ijl}, are chosen to be the normal with mean 0

and variance 0.15. Note that x_i is the highest concentration level for each dilution series. Thus even if x_i's are generated from a uniform distribution, the dilution series, $x_i - l$ are not uniformly distributed.

Figure 2.1 shows the distributions of D_{tot} based on 100 Monte Carlo data sets, using the Box-and-Whisker plot. The first panel corresponds to $k = 2$, the second panel corresponds to $k = 3$ and the third panel corresponds to $k = 4$. When $k = 3$ or 4, M-C and M-T lead to smaller maxima and smaller medians than M-S, and this can be more clearly seen in Table 2.1, where we present the summary statistics of D_{tot}. We see that M-S has a very large maximum, indicating occasional instability when the optimization is carried out in a higher dimensional parameter space. The results of M-C and M-T are generally better when $k = 3$ or 4 than when $k = 2$ although the difference is marginal.

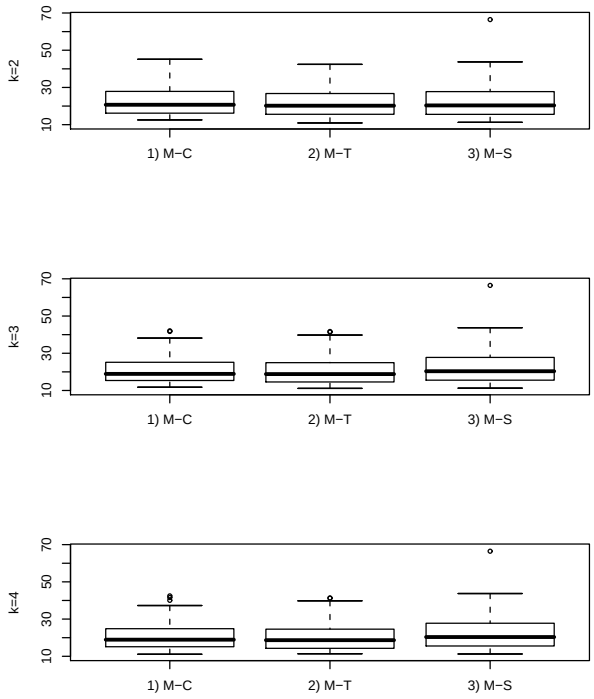

Figure 2.1: Box-and-Whisker plots of D_{tot} in Case 1. M-C and M-T are the multi-step procedures using component-wise minimization and trace minimization, respectively. M-S is the one-step least squares method.

We may examine the estimation result of the curve parameters, β_m, $m = 1, 2, 3$, based on 100 Monte Carlo data sets. Table 2.2 shows the summary statistics of 100 estimates. The result when $k = 4$ is given. The M-S method

Table 2.1: Summary statistics of D_{tot} in Case 1.

	$k=2$		$k=3$		$k=4$		
	M-C	M-T	M-C	M-T	M-C	M-T	M-S
Min	12.56	10.93	11.74	11.12	11.10	11.38	11.25
Median	20.69	20.19	18.92	18.81	18.93	18.67	20.35
Mean	22.24	21.68	20.70	20.40	20.54	20.37	22.48
Max	45.13	42.39	42.05	41.70	42.45	41.30	66.46

Table 2.2: Estimation result for the curve parameters in Case 1. The true value of $(\beta_1, \beta_2, \beta_3)$ is (10, 6, 0.5).

		M-C	M-T	M-S
β_1	Min	9.95	9.93	0.00
	Median	10.03	10.02	9.99
	Mean	10.03	10.02	9.28
	Max	10.13	10.13	10.15
	MSE	0.00	0.00	4.91
β_2	Min	5.81	5.82	5.86
	Median	5.94	5.97	6.03
	Mean	5.94	5.97	8.15
	Max	6.10	6.11	31.16
	MSE	0.01	0.01	33.71
β_3	Min	0.48	0.48	0.06
	Median	0.50	0.50	0.50
	Mean	0.50	0.50	0.42
	Max	0.52	0.53	0.52
	MSE	0.00	0.00	0.03

often produces very poor curve estimates. The mean squared error is also given in Table 2.2, and M-S produces large mean squared errors for all three parameters.

2.4.4 Simulation Case 2

In the second case, we generate data from the Sigmodial model with $(\beta_1, \beta_2, \beta_3) = (1, 1.5, 2)$. A total of 480 samples are used with $r = 3$ dilution series of length $l = 6$ for each sample. The error distribution is chosen to be normal with mean 0 and variance $(0.14)^2$. The true values of x_i are taken from the empirical distribution of the 96 estimated $\hat{x}_i$ values from the array $pThr308AKT$ used in [10].

Table 2.3: Summary statistics of D_{tot} in Case 2.

	$k=2$		$k=3$		$k=4$		
	M-C	M-T	M-C	M-T	M-C	M-T	M-S
Min	239.4	240.4	245.2	245.4	242.7	244.8	259.2
Median	264.4	264.9	266.0	267.3	262.1	264.1	280.5
Mean	268.9	268.9	270.0	271.4	267.1	268.3	284.6
Max	369.6	372.6	367.4	369.8	371.5	373.2	387.3

Figure 2.2 shows the the distributions of D_{tot} based on 500 Monte Carlo data sets. We note that M-C and M-T have lower values of D_{tot}, which can be more clearly seen in Table 2.3. If we compare data set by data set, the percentage of out-performance of M-C over M-S is 97.6%, and the percentage of out-performance of M-T over M-S is 98.2% in terms of D_{tot}.

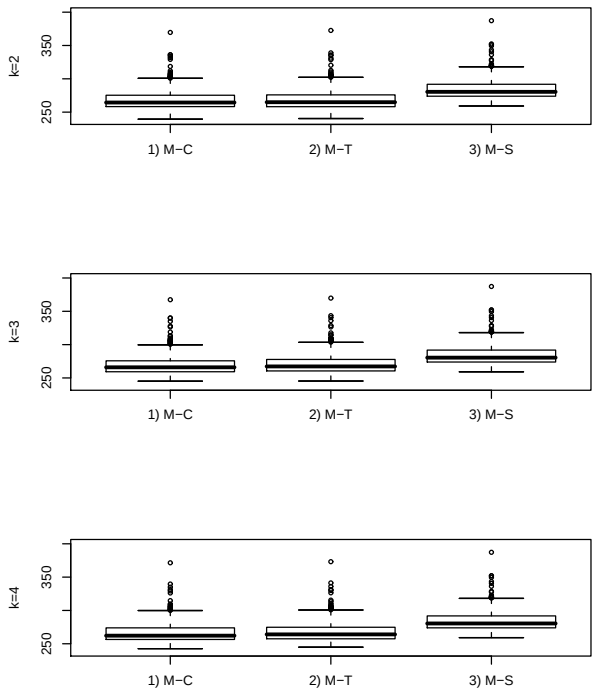

Figure 2.2: Box-and-Whisker plots of D_{tot} in Case 2.

Table 2.4 shows the summary statistics of 500 estimates for the curve parameters. For all three procedures, the parameters are very accurately estimated. The true values of $\beta_1, \beta_2, \beta_3$ are 1, 1.5, and 2, respectively.

Table 2.4: Estimation result for the curve parameters in Case 2. The true value of $(\beta_1, \beta_2, \beta_3)$ is (1, 1.5, 2).

		M-C	M-T	M-S
β_1	Min	0.99	0.99	0.99
	Median	1.00	1.00	1.00
	Mean	1.00	1.00	1.00
	Max	1.01	1.01	1.01
	MSE	0.00	0.00	0.00
β_2	Min	1.48	1.49	1.49
	Median	1.50	1.50	1.51
	Mean	1.50	1.50	1.50
	Max	1.52	1.52	1.52
	MSE	0.00	0.00	0.00
β_3	Min	1.87	1.87	1.93
	Median	1.94	1.96	2.00
	Mean	1.94	1.96	2.00
	Max	2.02	2.03	2.06
	MSE	0.00	0.00	0.00

2.4.5 Simulation Case 3

In the third case, we generate the concentration levels from a skew-normal distribution. From Figure 2.3, we find that the skew-normal distribution with location, scale, and shape parameters (3,2,-3) matches closely the empirical distribution of x_i estimates from real data analysis. The top two panels in Figure 2.3 show the empirical distributions of x_i estimates for the array of the protein $pThr308AKT$: log transformation of data is applied in the left panel and no transformation is applied in the right panel. We will discuss the transformation of data in details later in Section 2.5. A bump in the lower tail occurs due to the Winsorization. The bottom two panels show the empirical distributions of x_i estimates for the arrays of AKT and $pmTor$. The solid line in all panels represents the skew-normal distribution with location, scale, and shape parameters (3,2,-3). In this case, we use $\beta_1 = 1, \beta_2 = 1.5, \beta_3 = 1, s = 96, r = 3, l = 6$. The error terms are generated again from the normal distribution with mean zero and variance $(0.14)^2$.

Figure 2.4 shows the distributions of D_{tot} based on 500 Monte Carlo data sets. The M-C and M-T procedures (when $k = 4$) lead to smaller D_{tot} overall. It can be more easily seen in Table 2.5. If we compare case by case, M-C outperforms M-S in 60.2% of the time, and M-T outperforms M-S in 65.6% of the time in terms of D_{tot}.

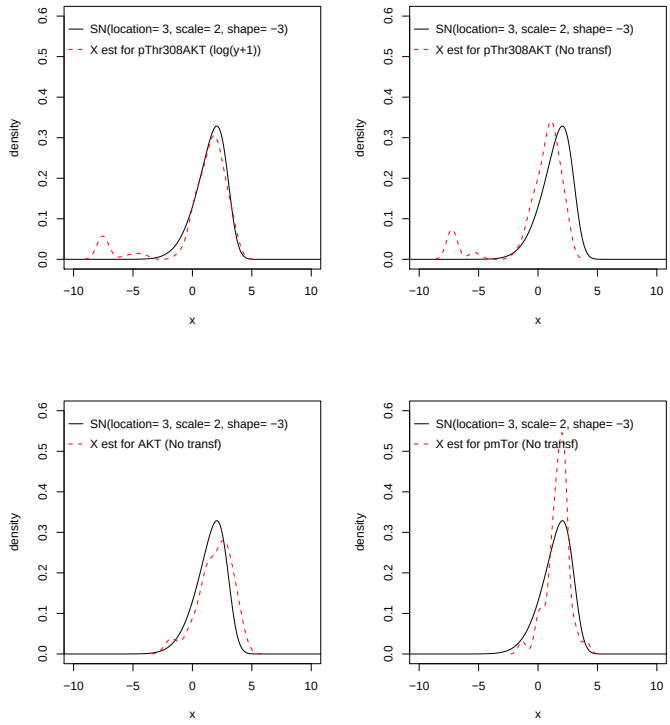

Figure 2.3: The empirical distribution of x_i estimates (the dashed line) versus the skew-normal distribution with location, scale, and shape parameters (3,2,-3) (the solid line). a) The empirical distributions of x_i estimates in the top two panels are based on the array of $pThr308AKT$ (log transformation of data is applied in the left panel and no transformation is applied in the right panel). b) The empirical distributions of x_i estimates in the bottom two panels are based on the arrays of AKT and $pmTor$.

Table 2.5: Summary statistics of D_{tot} in Case 3.

	$k=2$		$k=3$		$k=4$		
	M-C	M-T	M-C	M-T	M-C	M-T	M-S
Min	9.47	9.84	9.31	9.52	9.30	9.43	9.20
Median	15.65	15.87	15.14	14.91	14.96	14.72	15.78
Mean	16.66	17.06	16.25	16.05	16.09	15.89	17.07
Max	35.76	40.90	34.97	34.72	34.30	34.68	42.38

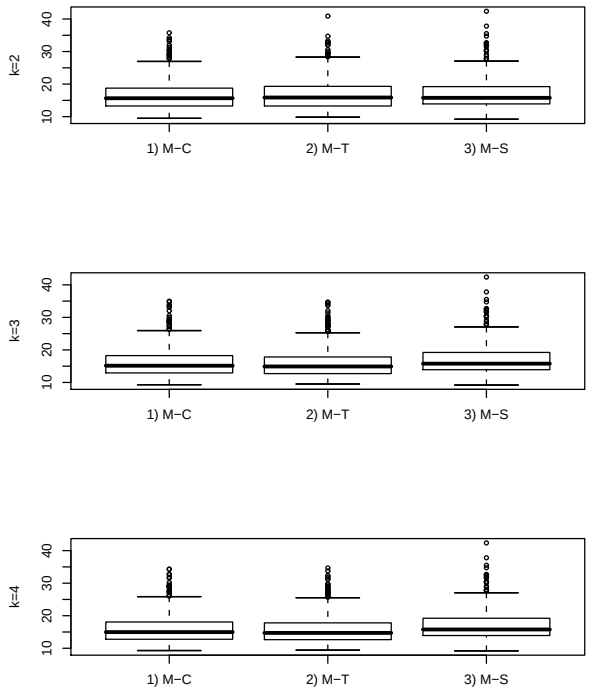

Figure 2.4: Box-and-Whisker plots of D_{tot} in Case 3.

Table 2.6: Estimation result for the curve parameters in Case 3. The true value of $(\beta_1, \beta_2, \beta_3)$ is (1, 1.5, 1).

		M-C	M-T	M-S
β_1	Min	0.98	0.89	0.64
	Median	1.01	1.00	0.84
	Mean	1.01	1.00	0.86
	Max	1.04	1.03	1.02
	MSE	0.00	0.00	0.03
β_2	Min	1.38	1.41	1.50
	Median	1.46	1.49	2.70
	Mean	1.46	1.50	2.49
	Max	1.54	1.83	3.30
	MSE	0.00	0.00	1.24
β_3	Min	0.90	0.55	0.39
	Median	0.98	0.98	0.49
	Mean	0.98	0.98	0.58
	Max	1.08	1.07	1.04
	MSE	0.00	0.00	0.21

Closer inspections show that the M-S method often produces poor curve estimates as shown in Table 2.6. On the other hand, M-C and M-T give quite accurate estimates for the curve parameters with very small mean squared errors.

2.4.6 Simulation Case 4

In this case, we generate data from the Sigmoidal model with $\beta_1 = 5, \beta_2 = 1.5, \beta_3 = 2$, and the error distribution of $N(0, 0.14^2)$. Each data set consists 96 samples, each of which consists of $r = 3$ dilution series of length $l = 6$. The values of x_i are generated from the skew-normal distribution with location, scale, and shape parameters (4,2,-2). We see in Figure 2.5 that this skew-normal distribution matches quite closely the empirical distribution of x_i estimates from real data analysis.

Figure 2.6 shows the the distributions of D_{tot} based on 500 Monte Carlo data sets. When $k = 4$, M-C and M-T have lower values of D_{tot} than M-S. If we compare case by case, M-C outperforms M-S in 75.8% of the time, and M-T outperforms M-S in 77.4% of the time in terms of D_{tot}. The performance of M-T significantly improves when we combine more samples into one group.

As for the curve parameters, M-S often gives very poor estimates as shown

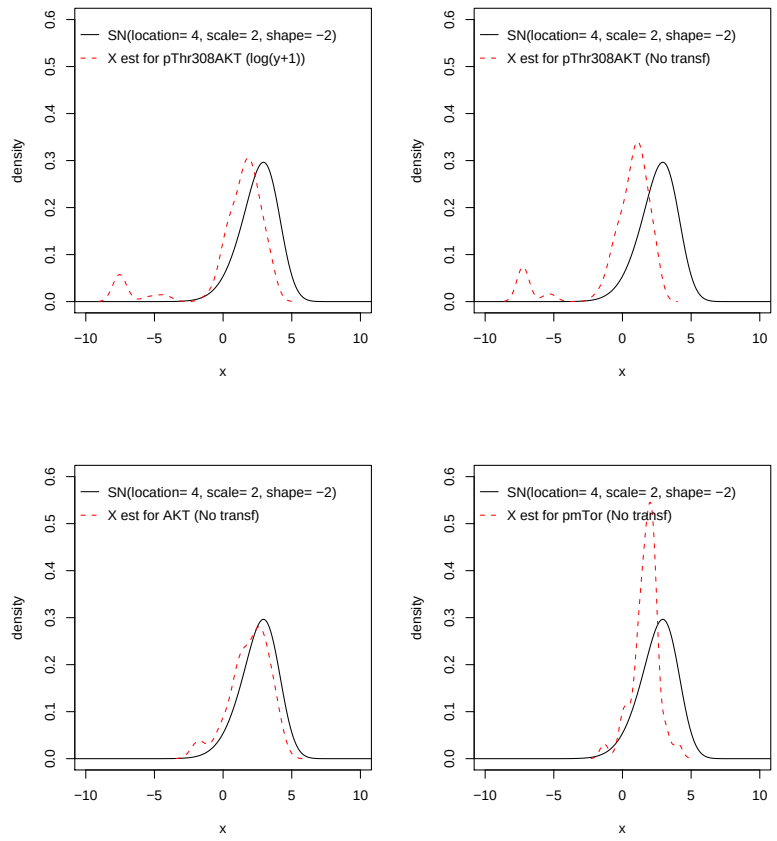

Figure 2.5: The empirical distribution of x_i estimates (the dashed line) versus the skew-normal distribution with location, scale, and shape parameters (4,2,-2) (the solid line). a) The empirical distributions of x_i estimates in the top two panels are based on the array of $pThr308AKT$ (log transformation of data is applied in the left panel and no transformation is applied in the right panel). b) The empirical distributions of x_i estimates in the bottom two panels are based on the arrays of AKT and $pmTor$.

Table 2.7: Summary statistics of D_{tot} in Case 4.

	$k=2$		$k=3$		$k=4$		
	M-C	M-T	M-C	M-T	M-C	M-T	M-S
Min	2.73	3.21	2.80	2.88	2.91	2.81	3.90
Median	10.39	10.43	9.53	9.38	8.52	8.31	34.64
Mean	11.29	17.60	10.44	12.54	9.57	10.16	34.12
Max	38.17	288.07	35.93	246.55	35.45	105.22	166.63

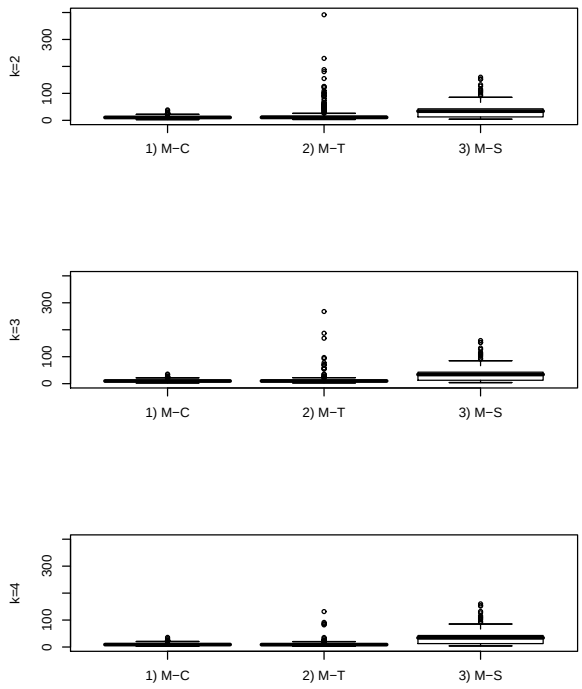

Figure 2.6: Box-and-Whisker plots of D_{tot} in Case 4.

Table 2.8: Estimation result for the curve parameters in Case 4. The true value of $(\beta_1, \beta_2, \beta_3)$ is (5, 1.5, 2).

		M-C	M-T	M-S
β_1	Min	4.98	4.88	0.00
	Median	5.00	5.00	3.98
	Mean	5.00	5.00	4.02
	Max	5.02	5.02	5.03
	MSE	0.00	0.00	1.92
β_2	Min	1.46	1.47	1.47
	Median	1.50	1.50	5.06
	Mean	1.50	1.50	4.49
	Max	1.54	1.72	13.23
	MSE	0.00	0.00	15.49
β_3	Min	1.83	0.54	0.08
	Median	1.96	1.97	0.25
	Mean	1.96	1.96	0.84
	Max	2.11	2.14	2.15
	MSE	0.00	0.02	2.07

in Table 2.8. On the other hand, our method (M-C and M-T) provides quite accurate estimates with very small mean squared errors.

2.4.7 Asymptotic Variance Estimates for Concentration Level Estimates

In order to assess the uncertainty of x_i estimate, we need to compute the variance of the estimate. A valid variance computation of x_i estimate is not given in [10]. In this section, we provide an asymptotic variance of x_i estimate using the asymptotic theory developed in Section 2.3. In addition, we evaluate the accuracy of the asymptotic variance estimates by comparing them with the Monte Carlo variances.

Monte Carlo Variances

We generate M Monte Carlo samples by randomly drawing the error terms from their true distribution, and for each Monte Carlo sample, we estimate x_i. The sample variance of M estimates for x_i, the Monte Carlo variance, will be used as a benchmark for evaluating the accuracy of the asymptotic variance estimate for x_i estimate. All Monte Carlo data sets share the identical set of x_i $(i = 1, \cdots, s)$, which is designed to make the comparison

among the Monte Carlo data sets possible.

Asymptotic Variance Estimates

By Theorem 1, we have

$$\widehat{var}(\tilde{x}_i) = \hat{\sigma}^2 \Bigg[r \sum_{l=0}^{t-1} \frac{(\hat{\beta}_2^{(c)})^2 (\hat{\beta}_3^{(c)})^2 e^{-2\hat{\beta}_3^{(c)}(\tilde{x}_i - l)}}{\left(1 + e^{-\hat{\beta}_3^{(c)}(\tilde{x}_i - l)}\right)^4} \Bigg]^{-1},$$

where $\hat{\beta}_1^{(c)}, \hat{\beta}_2^{(c)}$ and $\hat{\beta}_3^{(c)}$ are the pooled curve parameter estimates in Step 3, $\tilde{x_i}$ is the final concentration level estimate in Step 4 of our proposed method, and

$$\hat{\sigma}^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \sum_{l=0}^{t-1} \Big(y_{ijl} - \hat{\beta}_1^{(c)} - \frac{\hat{\beta}_2^{(c)}}{1 + e^{-\hat{\beta}_3^{(c)}(\tilde{x}_i - l)}} \Big)^2 \Big/ \big(srt - (3+s) \big).$$

Based on this result, we compute $\widehat{var}(\tilde{x}_i)$ for each Monte Carlo sample, and use the average of those variance estimates to see how close they are to the Monte Carlo variances.

Results

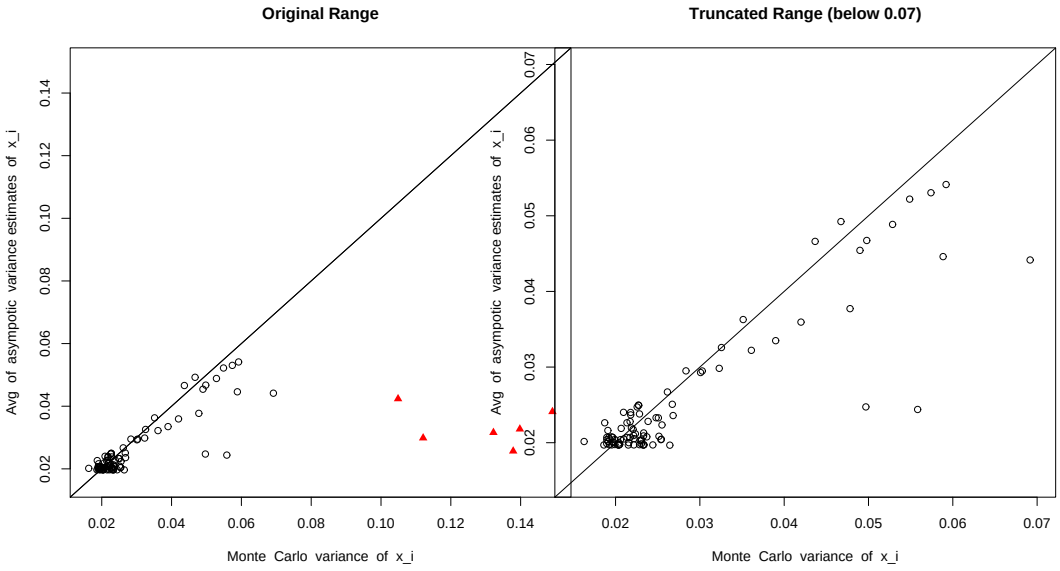

Figure 2.7: Monte Carlo variance versus average asymptotic variance estimate.

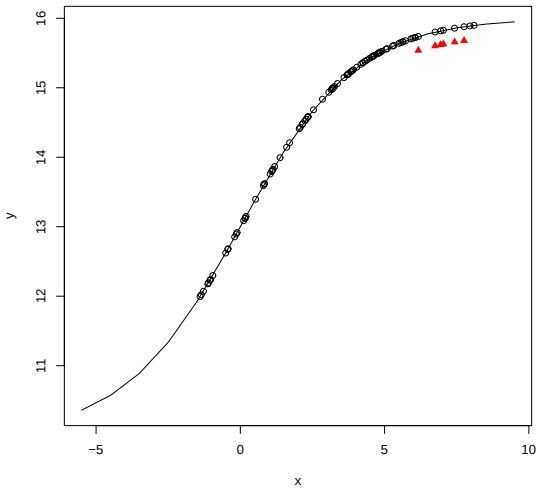

Figure 2.8: True curve and true x_i values: the points with the triangle symbol correspond to the six outliers in Figure 2.7.

The simulation setting of Case 1 is used when generating 100 Monte Carlo data sets. Figure 2.7 shows the average asymptotic variance estimate and the Monte Carlo variance of $\tilde{x}_i$ along with the 45-degree line. The first panel includes the variances for all $\tilde{x}_i'$s, and we note that for six $\tilde{x}_i'$s (represented with the triangle symbol), the asymptotic variance is underestimated. The second panel excludes these six outliers. Overall the average asymptotic variance estimates are close to the Monte Carlo variances with a few exceptions. The correlation coefficient between the average asymptotic variance estimates and the Monte Carlo variances for all the points is 0.54 and the correlation coefficient without six outlying points is 0.86. Figure 2.8 shows the true x_i on the true curve. The six outliers are again represented with the small triangles. To have a clearer look, the six points are shifted downwards slightly from the curve. As shown in Figure 2.8, the true x_i values for the six outliers are all near the tail area, which implies that an accurate estimation of the concentration level is more difficult in the tails than in the middle.

2.4.8 Confidence intervals

Finally, we consider a confidence interval for the relative concentration level using the bootstrap-t method ([19]). For each biological sample, we bootstrap r dilution series B times and compute a 100 $(1 - 2\alpha)$% confidence

interval on pairwise difference as follows:

$$CI_1 = \big(\hat{\eta}_i - \hat{se}(\hat{\eta}_i)\ \hat{q}^*_{sn}(1-\alpha),\ \hat{\eta}_i - \hat{se}(\hat{\eta}_i)\ \hat{q}^*_{sn}(\alpha)\big),$$

where $\hat{\eta}_i = \hat{x}_i - \hat{x}_{ref}$, and $\hat{q}^*_{sn}(\alpha)$ is the α^{th} sample quantile of the re-sampled statistics, $\{T^*_{sn,1}, ..., T^*_{sn,B}\}$ with $T^*_{sn} = (\hat{\eta}^*_i - \hat{\eta}_i)/se(\hat{\eta}^*_i)$, and $se(\hat{\eta}^*_i)$ is the standard error of $\hat{\eta}^*_i$ calculated from the bootstrap sample. We use the asymptotic result of Theorem 1 to obtain $\hat{se}(\hat{\eta}_i) = \widehat{var}(\hat{x}_i - \hat{x}_{ref})^{1/2}$, that is,

$$\widehat{var}(\hat{x}_i - \hat{x}_{ref}) = \hat{\sigma}^2\left[r\sum_{l=0}^{t-1}\frac{\hat{\beta}_2^2\hat{\beta}_3^2 e^{-2\hat{\beta}_3(\hat{x}_i - l)}}{\big(1+e^{-\hat{\beta}_3(\hat{x}_i-l)}\big)^4}\right]^{-1} + \hat{\sigma}^2\left[r\sum_{l=0}^{t-1}\frac{\hat{\beta}_2^2\hat{\beta}_3^2 e^{-2\hat{\beta}_3(\hat{x}_{ref} - l)}}{\big(1+e^{-\hat{\beta}_3(\hat{x}_{ref}-l)}\big)^4}\right]^{-1},$$

where the estimated values of β and x_i are used, and

$$\hat{\sigma}^2 = \sum_{i=1}^{s}\sum_{j=1}^{r}\sum_{l=0}^{t-1}\big(y_{ijl} - \hat{\beta}_1 - \frac{\hat{\beta}_2}{1+e^{-\hat{\beta}_3(\hat{x}_i - l)}}\big)^2 \Big/ \big(srt - (3+s)\big).$$

We shall denote as CI_1 the bootstrap confidence intervals described above. As shown in Lemma 1, the convergence rate of the curve parameter estimate is faster than that of the concentration estimate, and hence we may construct bootstrap confidence intervals by retaining the curve parameter estimate from the original data. This confidence interval, to be called CI_2, computes only the estimates of the concentration levels at each bootstrap sample, and therefore is much less computationally intensive than CI_1.

Based on the simulation setting of Case 3, we compute 90% bootstrap confidence intervals with 100 Monte Carlo data sets and the bootstrap size of $B = 100$. For each data set, we compute bootstrap confidence intervals for 95 pairwise differences $(x_i - x_{ref},\ i = 1, 2, ..., 96,\ i \neq ref)$, where the reference sample is taken to be the sample with the median concentration level. The coverage probability is the proportion of the intervals that contain the true values among all 95×100 intervals. In addition, we obtain the average interval length for each of the 95 pairwise differences. Using CI_2, the coverage probabilities of M-C, M-T, and M-S are 0.89, 0.89, and 0.88, respectively. As shown in Table 2.9, M-C and M-T have shorter confidence intervals than the M-S procedure on the average. Figure 2.9 displays the distribution of the interval length using the Box-and-Whisker plot. Overall, M-C and M-T result in shorter confidence intervals than M-S.

According to the result of Theorem 1, the concentration level estimates have the asymptotic normality, so we may consider an asymptotic confidence interval for the relative concentration level. Using the same Monte

Table 2.9: Coverage probabilities and average interval lengths based on simulation Case 3: the nominal level is 0.90.

		M-C	M-T	M-S
Bootstrap CI	Coverage prob	0.89	0.89	0.88
	Avg interval length	0.66	0.65	0.71
Asymptotic CI	Coverage prob	0.90	0.89	0.89
	Avg interval length	0.68	0.67	0.74

Carlo data sets as with the bootstrap confidence intervals, we compute 90% asymptotic confidence intervals for 95 pairwise differences. The coverage probabilities of M-C, M-T, and M-S are 0.90, 0.89, and 0.89, respectively. Again, as shown in Table 2.9, the average interval lengths of M-C and M-T are shorter than M-S. The asymptotic and the bootstrap confidence intervals produce very similar results.

For both bootstrap and asymptotic confidence intervals, two samples in the tails are excluded from computation.

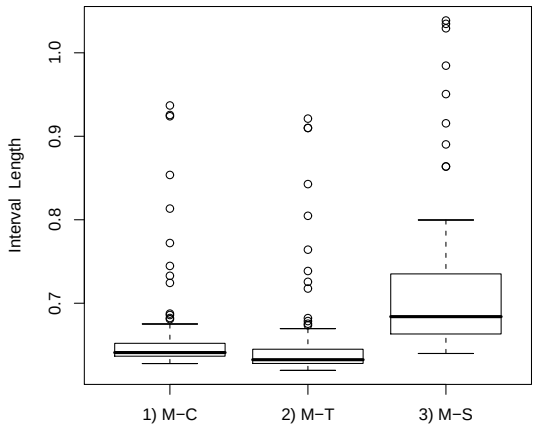

Figure 2.9: Box-and-Whisker plots of the bootstrap confidence interval length based on simulation Case 3.

2.5 Real Data Analysis

Now we present the results of real data analysis based on two different lysate arrays. The first array, the lysate array of the protein $pThr308AKT$ is mentioned in Section 3.4.1. In this array, 96 samples are 2-fold serially diluted 6 times and spotted on a slide in triplicate. Accordingly, there are $1728 = 96 \cdot 3 \cdot 6$ intensity measurements, but 15 measurements are not reliable because the spot intensities are lower than the background intensity and hence these measurements are not used for analysis. We apply both the proposed multi-step method (M-C with $k = 4$) and the existing least squares method (M-S) to this array.

Before that, we examine carefully the usefulness of data transformation. In Model (1.1), we assume the Sigmoidal relationship between the intensity level and the log-transformed concentration level. However, log transformation is often applied to the intensity level, too. In [10], log transformation is used for the intensity level and then a positive scalar is added to the log-transformed intensity level to ensure positive values. A statistical criterion for choosing an appropriate transformation is the comparison of the likelihoods. Suppose that we have

$$\begin{aligned} h(y_{ijl}) &= g(\underline{\beta}, x_i - l) + \epsilon_{ijl} \\ &= \beta_1 + \frac{\beta_2}{1 + e^{-\beta_3(x_i - l)}} + \epsilon_{ijl}, \end{aligned}$$

where y_{ijl} is the intensity level and x_i is the logarithm of the protein concentration level with base 2. A link function, h, is assumed to take one of the following forms:

$$h(y_{ijl}) = \begin{cases} y_{ijl} \\ log(y_{ijl} + c) \end{cases}$$

where c is positive number chosen to ensure positive values in all $h(y_{ijl})$. We assume that ϵ_{ijl} are independent and identically distributed ($i.i.d.$) following the normal distribution with mean zero and variance σ^2. Then the log-likelihood function is given by

$$l(\underline{\beta}, x_i, \sigma^2) = -\frac{srt}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{ijl} \big(y_{ijl} - g(\underline{\beta}, x_i - l)\big)^2,$$

if h is the identity function and

$$l(\underline{\beta}, x_i, \sigma^2) = -\frac{srt}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{ijl} \big(log(y_{ijl} + c) - g(\underline{\beta}, x_i - l)\big)^2 - \sum_{ijl} log(y_{ijl} + c),$$

if h is the log function.

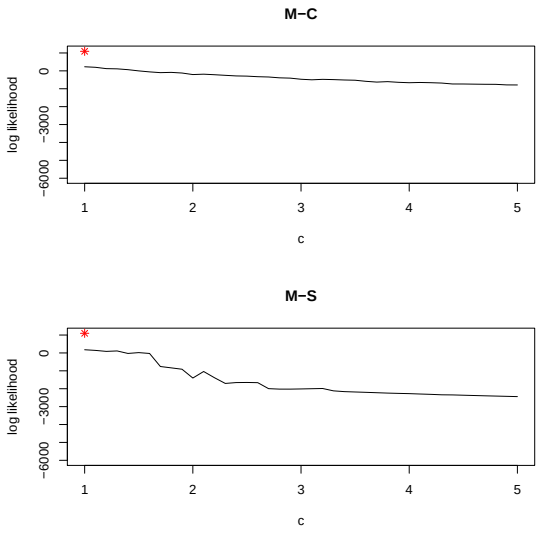

Figure 2.10: Log-likelihood function for $pThr308AKT$ array: the solid line and the star correspond to log transformation and no transformation, respectively.

Figure 2.10 plots the log-likelihood values computed with different transformations using M-C and M-S for the $pThr308AKT$ array. The solid line indicates the log-likelihood for the log transformation with $c \in [1, 5]$ and the star indicates the log-likelihood for the identity link, that is, no transformation. We see that no transformation has a higher log-likelihood than the log transformation does. The maximum of the log-likelihood with log transformation occurs at about $c = 1$ for both methods. We present the estimation results of no transformation and the log transformation with $c = 1$ in Figure 2.11. The left column corresponds to no transformation and the right column corresponds to the log transformation with $c = 1$. Top two rows display data points and fitted curves, where the x-axis is the median-centered concentration estimates. The fitted curves are adjusted accordingly to the median centering. Both methods fit the data reasonably well and give similar curve fits for either transformation. In fact, the curve parameter estimates are fairly close for two methods: with no transformation the estimates of M-C are $(0.11, 1.99, 0.67)$ and the estimates of M-S are $(0.11, 1.91, 0.75)$, while with the log transformation the estimates of M-C are $(0.12, 1.49, 0.59)$ and the estimates of M-S are $(0.15, 1.19, 0.92)$. The bottom row in Figure 2.11 compares the relative concentration level estimates using two methods for each transformation. The relative concentration estimate

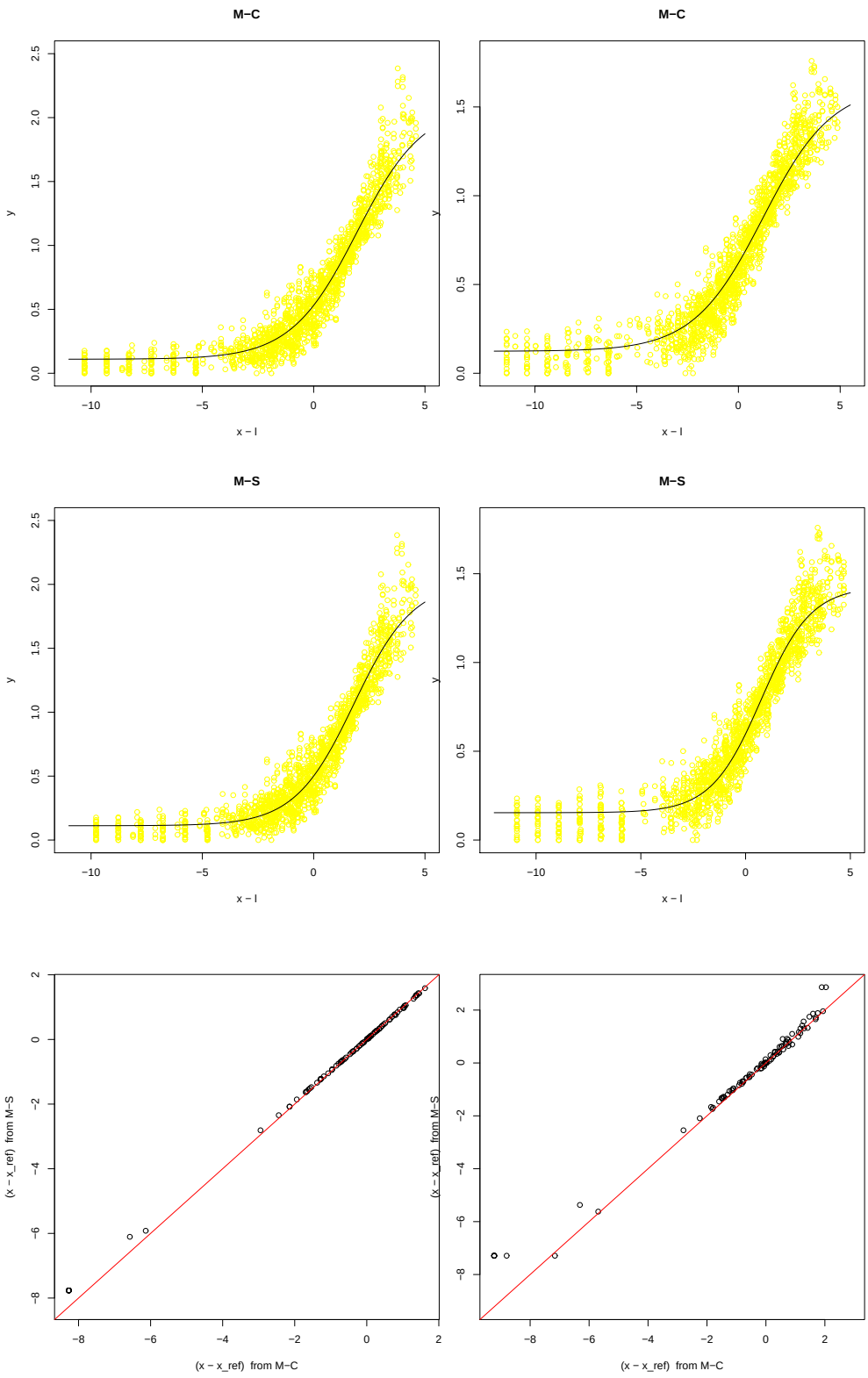

Figure 2.11: The left column corresponds to no transformation and the right column corresponds to the log transformation with $c = 1$ for $pThr308AKT$ array. Top two rows show data points and fitted curves and the bottom row compares the relative concentration level estimates using two methods.

is given by $\hat{x}_i - \hat{x}_{ref},\ i = 1, 2, ..., s$, where $\hat{x}_i$ and $\hat{x}_{ref}$ are the concentration estimates of the i^{th} sample and a reference sample, respectively. We use the middle sample based on the intensity measurements as the reference sample. We expect s dots to be close to the 45-degree line if the estimation results of two methods are similar, and indeed two methods yield the very similar estimates for both transformations as shown in this figure. It is not atypical that the proposed multi-step procedure and the least squares method on all samples give similar results, when both methods find the global optima in the nonlinear least squares.

In the lysate array of the protein AKT, 40 samples are diluted 2-fold eight times and spotted on a slide in duplicate and all the measurements are used for analysis. A detailed layout of this array can be found in [13]. Figure 2.12 shows the results based on AKT array. We first see that no transformation leads to a substantially larger log-likelihood than log transformation does. Therefore, we only present the result of no transformation and as shown in Figure 2.12, M-C and M-S fit the data very well and give similar fitted curves and concentration estimates. The curve parameter estimates are (472.06, 21801.42, 0.58) for M-C and (239.60, 21122.32, 0.63) for M-S.

Next, we compute the confidence intervals for the relative concentration levels using the bootstrap methods described in the previous section. The bootstrap size is 400, and the confidence level is chosen to be 90%. We construct two types of the bootstrap confidence intervals (CI_1 and CI_2) for $x_i - x_{ref}, i = 1, ..., 96, i \neq ref$, using M-C and M-S, and then assess how often the different procedures lead to the same conclusion regarding the signs of $x_i - x_{ref}$. Since the number of replicates is too small in AKT array ($r = 2$), we use the $pThr308AKT$ array. The 3×3 contingency tables in Table 2.10 cross classify 95 intervals by the signs of the intervals that are obtained under different procedures. The sign has three levels depending on whether the interval contains only negative values ($-$), or contains zero (Not significant), or only positive values ($+$). The first contingency table compares CI_1 using M-C with CI_1 using M-S. They agree very well and no opposite signs are seen from CI_1 and CI_2. The second and the third contingency tables compare CI_1 and CI_2 within each method. As we see, CI_1 and CI_2 agree perfectly for both M-C and M-S. Table 2.11 shows the bootstrap confidence intervals corresponding to the off-diagonal entries in the first contingency table in Table 2.10. We do not see considerable differences in those intervals even when the signs do not match exactly. With real data, we cannot be sure about the coverage probabilities of those confidence intervals. To see the interval length, we use Figure 2.13, where we display the distribution of the

interval length using the Box-and-Whisker plot. Overall, the M-C procedure results in shorter confidence intervals than M-S, which is consistent with what we learned from the simulation study in the previous section. Ten samples in the tails are excluded from this figure.

2.6 Conclusion

In this article, we consider a modified multi-step procedure for lysate array quantification based on the Sigmoidal model. Our modification from the least squares method of [10] simplifies both theory and computations. In theory, the modification enables us to verify asymptotic normality of the protein concentration estimates as the number of biological samples and the number of measurements per sample grow, thus providing a theoretical foundation for statistical inference. In computation, the modification employs the nonlinear optimization in lower parameter spaces, reducing the risk of being trapped in local minima. For most data sets, the least squares method and our proposed modification produce very similar results, but the advantages of a simpler asymptotic theory and better numerical stability in some of the data sets make the modification worthwhile. The proposed method will be more valuable in applications where the total number of samples, s, on the protein array is large.

We have focused on the specific Sigmoidal model in the paper with *i.i.d.* errors. The basic principle of grouping and pooling in the multi-step procedure applies readily to other models. Relaxation of the *i.i.d.* errors assumption is quite important. One attempt to allow for *non-i.i.d.* errors is given in the next chapter.

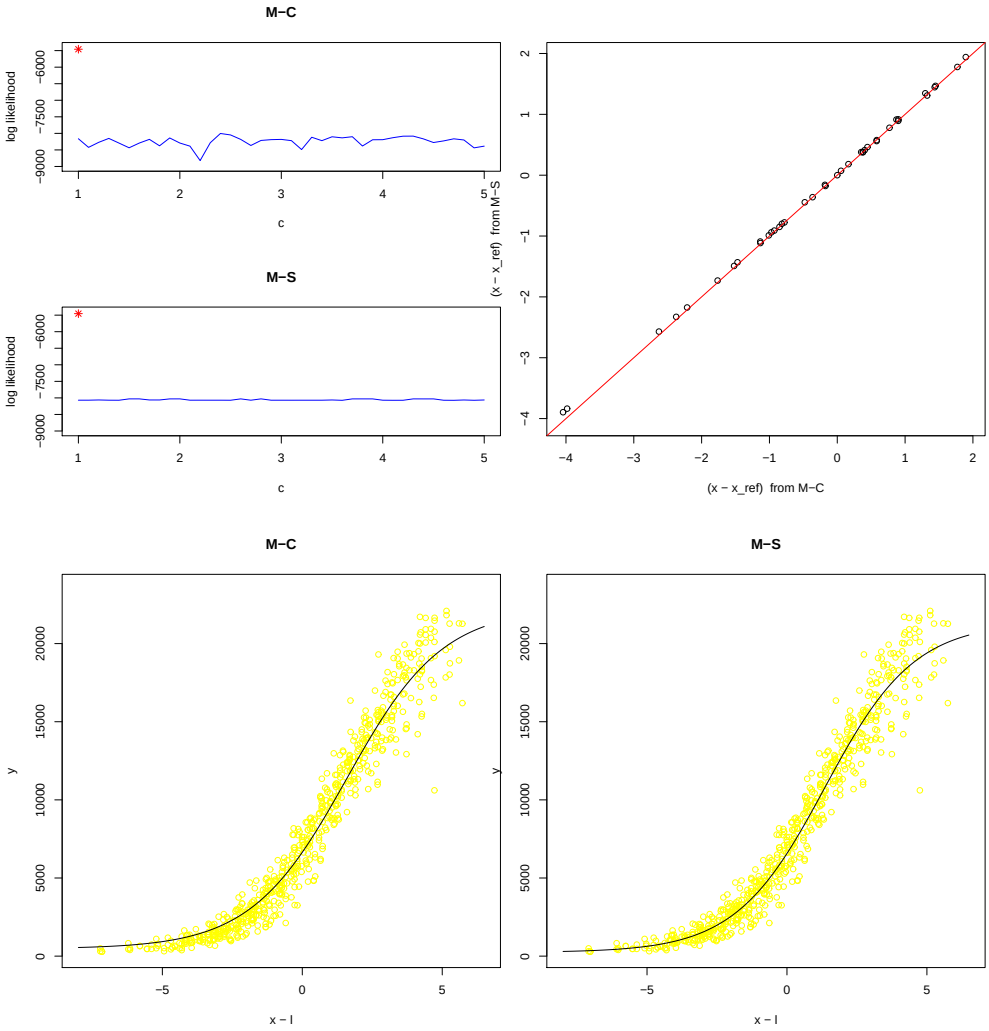

Figure 2.12: The upper left corner corresponds to the log-likelihood function (the solid line is for log transformation and the star is for no transformation), the upper right corner corresponds to the comparison of the relative concentration level estimates, and the bottom row shows data points and fitted curves using two methods (without transformation) for AKT array.

Table 2.10: Contingency table categorized by the signs of the intervals using two different ways for $pThr308AKT$.

	CI_1 using M-S			
CI_1 using M-C	$-$	Not significant	$+$	Total
$-$	34	3	0	37
Not significant	0	18	0	18
$+$	0	0	40	40
Total	34	21	40	95
	CI_2 using M-C			
CI_1 using M-C	$-$	Not significant	$+$	Total
$-$	37	0	0	37
Not significant	0	18	0	18
$+$	0	0	40	40
Total	37	18	40	95
	CI_2 using M-S			
CI_1 using M-S	$-$	Not significant	$+$	Total
$-$	34	0	0	34
Not significant	0	21	0	21
$+$	0	0	40	40
Total	34	21	40	95

Table 2.11: Bootstrap confidence intervals for $x_i - x_{ref}$ that correspond to the off-diagonal entries in the first contingency table in Table 2.10.

CI_1 using M-C	CI_1 using M-S
$(-0.68, -0.03)$	$(-0.65, 0.02)$
$(-0.69, -0.03)$	$(-0.40, 0.28)$
$(-1.16, -0.50)$	$(-0.59, 0.15)$

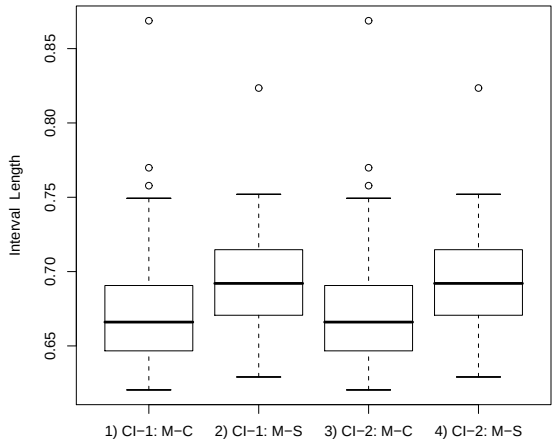

Figure 2.13: Box-and-Whisker plots of the confidence interval length for $pThr308AKT$. The first two columns are for the confidence intervals CI_1 under M-C and M-S procedures, and the last two columns are for the confidence intervals CI_2.

CHAPTER 3

PROTEIN LYSATE ARRAY QUANTIFICATION METHOD UNDER *NON-I.I.D.* MODEL

3.1 Introduction

Previous studies of protein array quantification including [10] are based on the assumption that the errors are independent. However, when we examine the residuals obtained under the *i.i.d.* error assumption, this assumption appears very questionable. The nature of the experiment warrants the existence of correlation, too. The repeated measurements of each biological sample are likely to be correlated. Also the measurements within dilution series come from the same replicate of the same biological sample, and are thereby likely to have correlation. In this chapter, we consider the complexity of the correlation structure, and introduce a new model that can allow for the dependence structure of the errors by a nonlinear mixed effects model. Based on the new model, we consider a method to approximate the joint maximum likelihood estimator of all the parameters. We show empirically that the new model is valuable for the protein lysate array quantification. In future work, we will employ two other methods that may improve computational efficiency and stability. The first is the EM algorithm and the second is a Bayesian approach with MCMC. In addition, we will develop an asymptotic theory on the joint maximum likelihood estimator based on the new model.

The rest of this chapter is organized as follows. In Section 3.2, we raise concerns about the *i.i.d.* error assumption based on the examination of real data, and introduce a model that allows for the dependence structure. In Section 3.3, we propose a method to approximate the joint maximum likelihood estimator. As preliminary studies, simulated data and real data are analyzed in this section. In Section 3.4, we perform several simulation studies to evaluate the performance of the proposed method, in comparison with the existing method. Future work is discussed in the same section.

3.2 Non-IID Model

Most prior work on protein array quantification is based on the assumption that the errors are independent. The single-step least squares procedure of [10] that is many times cited in the literature works when the errors are $i.i.d.$ In this section, we raise concerns about this assumption based on the examination of real data, and introduce a model that allows for dependence in the data.

3.2.1 Motivation

We examine the empirical residuals after applying the single-step least squares procedure discussed in [10] on two arrays for the protein AKT and the protein $pmTor$. These arrays were produced at M.D. Anderson and for both arrays, 40 biological samples are 2-fold serially diluted eight times and spotted on a slide in duplicate. A detailed layout is described in [13].

A statistical criterion for measuring the correlation between two random vectors is canonical correlation, which is the maximum correlation between linear combinations of the two vectors ([20]). If we treat the residuals from a dilution series as a random vector, then canonical correlation between two sets of the residuals from two dilution series of a certain biological sample can be thought of as the correlation between two replicates within the sample. For both arrays, we computed the canonical correlations between two dilution series and we obtained 0.92 for AKT array, 0.96 for $pmTor$ array. They indicate that the replicates within the sample may be highly correlated. In addition, Figure 3.1 shows that the residuals from two dilution series are highly linearly correlated. The first panel presents the residuals of 40 samples at the first dilution level for AKT array, and the second panel presents the residuals of 40 samples at the fourth dilution level for $pmTor$ array.

These findings lead us to consider a new model that allows for $non\text{-}i.i.d.$ error structures, which will be discussed in the next section.

3.2.2 Mixed Effects Model

In this section, we introduce a model that takes into account a possible dependence structure of the lysate array data. We assume the Sigmoidal curve for the relationship between the concentration level and the intensity measurements. However, unlike in Chapter 2, we assume that there are

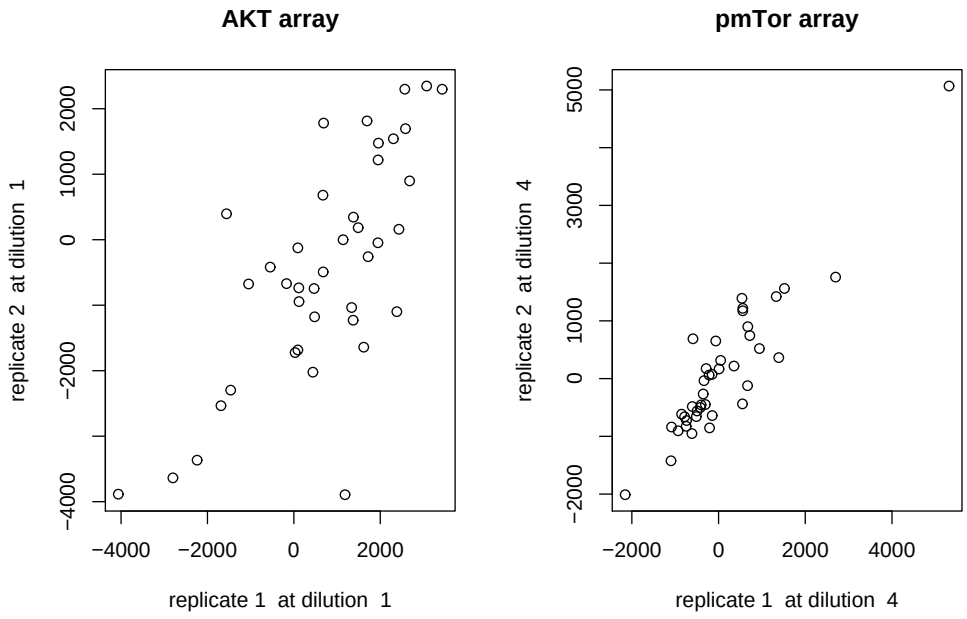

Figure 3.1: The residuals from two dilution series of 40 samples for AKT array and $pmTor$ array.

random effects of the dilution series nested within sample. In addition, we assume heteroscedastic measurement errors for different dilution levels. Consider the following model:

$$y_{ijl} = g(\underline{\beta}, x_i - l) + \delta_{ij} + \epsilon_{ijl}, \tag{3.1}$$

where $i = 1, ..., s,\ \ j = 1, ..., r,\ \ l = 0, ..., (t-1),\ \ y_{ijl}$ is the intensity measurement at the l^{th} dilution of the j^{th} replicate for the i^{th} sample, x_i is the logarithm of the protein concentration of the i^{th} sample, and $g(\underline{\beta}, x_i - l) = \beta_1 + \frac{\beta_2}{1+e^{-\beta_3(x_i-l)}}$. We use the notation, δ_{ij} for the random effect of the j^{th} replicate, nested within the i sample, but assumed to be independent for different samples. We assume that the random effects, δ_{ij}, are independent of the errors, ϵ_{ijl}. In addition, δ_{ij} are assumed to be correlated for different replicates. Assume that $\underline{\delta}_i = (\delta_{i1}, \cdots, \delta_{ir})'$ follows the normal distribution:

$$\underline{\delta}_i \sim N_r\big(\underline{0},\ V_\delta\big),$$

where V_δ is an $r \times r$ matrix taking the form,

$$V_\delta = \sigma_\delta^2 \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ & & \dots & \\ \rho & \rho & \dots & 1 \end{pmatrix}.$$

For the error terms, ϵ_{ijl}, we assume that they are mutually independent and $\underline{\epsilon}_{ij} = (\epsilon_{ij0}, ..., \epsilon_{ij(t-1)})'$ follows the normal distribution:

$$\underline{\epsilon}_{ij} \sim N_t\big(\underline{0},\ V_\epsilon\big),$$

where V_ϵ is a $t \times t$ matrix taking the form,

$$V_\epsilon = \begin{pmatrix} \sigma_{\epsilon_0}^2 & 0 & \dots & 0 \\ 0 & \sigma_{\epsilon_1}^2 & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & \sigma_{\epsilon_{(t-1)}}^2 \end{pmatrix}.$$

Now we denote the vector of all measurements by

$$\begin{aligned} \underline{Y} &= \Big(y_{ijl}\Big)_{i=1,...,s,\ j=1,...,r,\ l=0,...,(t-1)} \\ &= \Big(\underline{y}_1{}', \underline{y}_2{}', \cdots, \underline{y}_s{}'\Big)', \end{aligned}$$

where $\underline{y}_i = \big(y_{i10}, \dots, y_{i1(t-1)}, y_{i20}, \dots, y_{i2(t-1)}, \dots, y_{ir0}, \dots, y_{ir(t-1)}\big)'$ is the vector of the measurements of length rt, stacked first by replicate and then by dilution level for the i^{th} sample. Also let $\underline{G}$, $\underline{\Delta}$, and $\underline{\Omega}$ denote the vectors of length srt for the mean effects, the random effects for replicate, and the within-group errors, respectively. These vectors are stacked by sample, replicate and sample so that the first rt observations correspond to the first sample, the first t observations correspond to the first replicate of the first sample, and so on. Then we have

$$\underline{Y} = \underline{G} + \underline{\Delta} + \underline{\Omega},$$

and $\underline{\Upsilon} = \underline{\Delta} + \underline{\Omega}$ follows the normal distribution with $E(\underline{\Upsilon}) = \underline{0}$ and $cov(\underline{\Upsilon}) = I_s \otimes \big(V_\delta \otimes J_t + I_r \otimes V_\epsilon\big)$, where $J_r = \underline{1}_r \underline{1}_r'$ and $\underline{1}_r = (1, \cdots, 1)'$. The following conditional (co)variances of the measurements given the mean effect, g, may

help us to understand better the dependence structure,

$$\begin{aligned} var(y_{ijl} \mid g) &= \sigma_\delta^2 + \sigma_{\epsilon_l}^2, \\ cov(y_{ijl}, y_{ijl^*} \mid g) &= \sigma_\delta^2, \;\; l \neq l^*, \\ cov(y_{ij\cdot}, y_{ij^*\cdot} \mid g) &= \rho\sigma_\delta^2, \;\; j \neq j^*, \\ cov(y_{i\cdot\cdot}, y_{i^*\cdot\cdot} \mid g) &= 0, \;\; i \neq i^*. \end{aligned} \tag{3.2}$$

In order for the variance-covariance matrix of $\underline{\Upsilon}$ to be positive definite, we need the following conditions:

$$\sigma_\delta^2 > 0, \quad \sigma_{\epsilon_l}^2 > 0, \quad -\frac{1}{r-1} < \rho < 1. \tag{3.3}$$

3.3 Estimation Methods

To show Model (3.1) adds value to the analysis, we propose a method that approximates the joint maximum likelihood estimate (this method, hereafter, will be referred to as the joint maximum likelihood estimation). As preliminary studies, we evaluate how well the joint MLE works for estimating the variance components by using simulated data and real data. Besides the joint MLE, another method, two-stage estimation, is considered but it turns out that this method often fails to accurately estimate the variance components.

3.3.1 Joint Maximum Likelihood Estimation

We consider the first method for estimating the parameters in Model (3.1). Since it approximates the joint maximum likelihood estimate, we will refer this method as the joint maximum likelihood estimation. In this method, we estimate jointly all the parameters that maximize the likelihood function, in an iterative way: estimating x_i, V_δ and V_ϵ given $\underline{\beta}$ and estimating $\underline{\beta}$ given x_i, V_δ and V_ϵ.

When estimating x_i, V_δ and V_ϵ given $\underline{\beta}$, we combine h samples, which allows us to use the measurements of several samples for the estimation and thus to attain more accurate and stable estimates than the estimates from a single sample. However, if h is too large (e.g. $h = s$), the estimation may suffer from a high-dimensional parameter space. On the other hand, when estimating $\underline{\beta}$ given x_i, V_δ and V_ϵ, we use the measurements from all samples. After that, to obtain more refined and stable estimates, we may

use further estimation steps. The details are given below.

Step 1: Estimate $x_i,\ \sigma_\delta^2,\ \rho$, and $\sigma_{\epsilon_l}^2$ given $\underline{\beta}$.

(a) Choose a small value of h and combine h samples into one group, resulting in s/h groups (or its integer part).

(b) For each group, find the estimates of x_i, σ_δ^2, ρ and $\sigma_{\epsilon_l}^2$ that minimize the following function with the constraints of (3.3):

$$-lnL\big(x_i, \sigma_\delta^2, \rho, \sigma_{\epsilon_l}^2 \mid y_{ijl}, \hat{\underline{\beta}}\big) \propto h\ ln\big(|Q|\big) + \sum_{i=1}^{h} \big(\underline{y}_i - g_i\big)' Q^{-1} \big(\underline{y}_i - g_i\big),$$

where $i = 1, \cdots, h,\ j = 1, \cdots, r,\ l = 0, \cdots, (t-1),\ Q = V_\delta \otimes J_t + I_r \otimes V_\epsilon,\ \underline{y}_i$ is the vector for the measurements of the i^{th} sample as defined in Section 3.2.2, and g_i is the rt by 1 vector for the mean effects of the i^{th} sample with $\underline{\beta}$ being fixed as $\hat{\underline{\beta}}$, obtained from the single-step procedure of [10]. Hence, g_i consists of r identical vectors of length t, which is given by

$$\Big(g\big(\hat{\underline{\beta}},\ x_i\big),\ g\big(\hat{\underline{\beta}},\ x_i - 1\big), \cdots,\ g\big(\hat{\underline{\beta}},\ x_i - (t-1)\big)\Big)'.$$

(c) Find the median of the s/h estimates for each of σ_δ^2, ρ and $\sigma_{\epsilon_l}^2$, and construct the variance-covariance matrix based on those medians. The variance-covariance matrix is $\hat{Q} = V_\delta \otimes J_t + I_r \otimes V_\epsilon$.

Step 2: Estimate $\underline{\beta}$ given $x_i,\ \sigma_\delta^2,\ \rho$, and $\sigma_{\epsilon_l}^2$ by finding the estimate that minimizes the following function:

$$-lnL\big(\underline{\beta} \mid y_{ijl}, \hat{x}_i, \hat{Q}\big) \propto \sum_{i=1}^{s} \big(\underline{y}_i - \hat{g}_i\big)' (\hat{Q})^{-1} \big(\underline{y}_i - \hat{g}_i\big),$$

where $i = 1, \cdots, s,\ j = 1, \cdots, r,\ l = 0, \cdots, (t-1),\ \hat{x}_i$ and $\hat{Q}$ are the estimates obtained from Step 1, and $\hat{g}_i$ is the vector for the mean effects of the i^{th} sample with x_i being fixed as $\hat{x}_i$.

Step 3: Repeat Step 1 using the $\underline{\beta}$ estimate obtained from Step 2.

After this one-step iteration, we use one further estimation step to obtain more refined and stable estimates.

Step 4: Estimate σ_δ^2, ρ, and $\sigma_{\epsilon_l}^2$ using a linear mixed effects model, and estimate x_i and $\underline{\beta}$ using the multi-step procedure of Chapter 2.

(a) Find the the maximum likelihood estimates for $\sigma_\delta^2,\ \rho,$ and $\sigma_{\epsilon_l}^2$, in V_δ and V_ϵ, using the following linear mixed effects model as an approximate,

$$\hat{\eta}_{ijl} = \delta_{ij} + \epsilon_{ijl},$$

where $i = 1, \cdots, s,\ j = 1, \cdots, r,\ l = 0, \cdots, (t-1),\ \hat{\eta}_{ijl}$ are the residuals obtained with the most updated estimates for $\underline{\beta}$ and x_i, and δ_{ij} and ϵ_{ijl} are the random effects of the dilution series and the within-group errors, respectively, as defined in Model (3.1).

(b) Find the estimates for x_i and $\underline{\beta}$ using the multi-step procedure after a modification to implement the weighted least squares estimation method.

i. Find group-based estimates for x_i and $\underline{\beta}$ that minimize the following objective function:

$$\sum_{i=1}^{k} \big(\underline{y}_i - g_i\big)' \tilde{Q}^{-1} \big(\underline{y}_i - g_i\big),$$

where $\tilde{Q}$ is the variance-covariance matrix constructed with the most updated estimates for the variance components, $\underline{y}_i$ is the vector for the measurements of the i^{th} sample, and g_i is the vector for the mean effects of the i^{th} sample.

ii. Pool the group-based estimates for $\underline{\beta}$ using the same criteria as in Section 2.2.2.

iii. Given $\underline{\beta}$, estimate x_i that minimizes the following objective function:

$$\big(\underline{y}_i - \tilde{g}_i\big)' \tilde{Q}^{-1} \big(\underline{y}_i - \tilde{g}_i\big),$$

where $i = 1, \cdots, s,$ and $\tilde{g}_i$ is the vector for the mean effects of the i^{th} sample with β being fixed as the pooled estimate.

In our implementation, we use the R function *optim* for Steps 1 – 3 and Step 4 (b), and *lme* for Step 4 (a). In Step 1, the starting value of x_i is obtained from the single-step procedure, and for the starting values for σ_δ^2, ρ, and $\sigma_{\epsilon_l}^2$, we compute the sample moments of the first three (co)variances in (3.2), and then find the corresponding sample moments of σ_δ^2, $\sigma_{\epsilon_l}^2$, and ρ.

The joint maximum likelihood estimate of all the parameters could be obtained at once by maximizing one log-likelihood function, but the proposed one-step iteration reduces numerical challenges in optimization, because we avoid working directly with a high-dimensional parameter space.

3.3.2 Two-Stage Estimation

We consider the second method for estimating the parameters in Model (3.1). In this method, we first estimate x_i and $\underline{\beta}$ under the *i.i.d.* error assumption as in Chapter 2, and obtain the residuals. After that, we estimate all parameters using the method described in Step 4 of the joint MLE. Unlike the joint MLE, the variance components (V_δ and V_ϵ) are not estimated jointly with x_i and $\underline{\beta}$ in this method. By using simulated data, we will show in Section 3.3.3 that this separate estimation often produces biased estimates for the variance components. The details of the two-stage method are as follows.

Step 1: Find the estimates for x_i and $\underline{\beta}$ using the single-step least squares method discussed in [10], and then obtain the residuals, $\check{\eta}_{ijl}$.

Step 2: Estimate $\sigma^2_\delta,\ \rho,$ and $\sigma^2_{\epsilon_l}$ using a linear mixed effects model, and estimate x_i and $\underline{\beta}$ using the multi-step procedure of Chapter 2.

(a) Find the the maximum likelihood estimates for $\sigma^2_\delta,\ \rho,$ and $\sigma^2_{\epsilon_l}$, in V_δ and V_ϵ, using the following linear mixed effects model as an approximate,

$$\check{\eta}_{ijl} = \delta_{ij} + \epsilon_{ijl},$$

where $i = 1, \cdots, s,\ j = 1, \cdots, r,\ l = 0, \cdots, (t-1),$ $\check{\eta}_{ijl}$ are the residuals obtained by using the estimates for x_i and $\underline{\beta}$ from the previous step, and δ_{ij} and ϵ_{ijl} are the random effects of the dilution series and the within-group errors, respectively, as defined in Model (3.1).

(b) Find the estimates for x_i and $\underline{\beta}$ using the multi-step procedure after a modification to implement the weighted least squares estimation method.

i. Find group-based estimates for x_i and $\underline{\beta}$ that minimize the

following objective function:

$$\sum_{i=1}^{k} \big(\underline{y}_i - g_i\big)' \check{Q}^{-1} \big(\underline{y}_i - g_i\big),$$

where $\check{Q}$ is the variance-covariance matrix constructed by using the most updated estimates for the variance components.

ii. Pool the group-based estimates for $\underline{\beta}$ using the same criteria as in Section 2.2.2.

iii. Given $\underline{\beta}$, estimate x_i that minimizes the following objective function:

$$\big(\underline{y}_i - \check{g}_i\big)' \check{Q}^{-1} \big(\underline{y}_i - \check{g}_i\big),$$

where $i = 1, \cdots, s$, and $\check{g}_i$ is the vector for the mean effects of the i^{th} sample with β being fixed as the pooled estimate.

The main purpose of this method is that we hope to capture the real error structure by using the residuals obtained under the *i.i.d.* working assumption. However, it turns out that the residuals do not represent the true errors very well in terms of correlation, as we will show later in Section 3.3.3. One possible explanation is as follows. Suppose that three random variables, X_1, X_2 and X_3, are positively correlated. Yet, $X_1 - \bar{X}$ and $X_2 - \bar{X}$ are not necessarily positively correlated. This can happen especially when the number of $X's$ is small. For example, let $var(X_1) = var(X_2) = var(X_3) = 1$ and $cov(X_1, X_2) = cov(X_1, X_3) = cov(X_2, X_3) = \varrho > 0$. It follows that

$$\begin{aligned} cov(X_1 - \bar{X},\ X_2 - \bar{X}) &= \Big(\frac{2}{3}, -\frac{1}{3}, -\frac{1}{3}\Big)\, cov(X_1, X_2)\, \Big(-\frac{1}{3}, \frac{2}{3}, -\frac{1}{3}\Big)' \\ &= -\frac{1}{3}(1 - \varrho) \leq 0. \end{aligned}$$

In a similar way, even if two dilution series of a certain biological sample are correlated, their residuals might not have been correlated, especially when the number of replicates $(= r)$ is small.

3.3.3 Studies with Simulated Data

In this section, we evaluate how well two methods work for estimating the variance components in Model (3.1) using simulated data. We consider mainly two scenarios:

(i) when σ_δ^2 is strictly positive (when the random effects exist), and

(ii) when σ_δ^2 equals zero (when the random effects do not exist).

If there are no random effects, the measurements can be considered as independent. For each scenario, we assume either

(i) heteroscedastic measurement errors (the different diagonal elements in V_ϵ) or

(ii) homoscedastic measurement errors (the common diagonal element in V_ϵ).

If $\sigma_\delta^2 = 0$ and V_ϵ has the common diagonal element, then the measurements are considered as $i.i.d.$ We use the following values: $s = 96,\ r = 3,\ t = 6$, and $\underline{\beta} = (10, 15, 1)'$. We generate the concentration levels from the skew-normal distribution with location, scale, and shape parameters (3,2,-3), whose density matches closely the empirical distribution of x_i estimates from the real data analyses of array $pThr308AKT$, array AKT, and array $pmTor$ (see Figure 2.3). For the joint MLE, we combine $h = 12$ samples when estimating x_i, V_δ and V_ϵ. In the first scenario, we assume that

$$\sigma_\delta^2 = 1,\ \ \rho = 0.7,\ \ (\sigma_{\epsilon_0}^2, \sigma_{\epsilon_1}^2, \sigma_{\epsilon_2}^2, \sigma_{\epsilon_3}^2, \sigma_{\epsilon_4}^2, \sigma_{\epsilon_5}^2)' = (0.40, 0.35, 0.30, 0.25, 0.20, 0.15)'.$$

The estimation results from two methods are given in Table 3.1. The two-stage estimation gives very poor estimates overall. Especially the estimate for σ_δ^2 is too small and the estimate for ρ is even negative although the true correlation coefficient is quite high at 0.7. On the other hand, the estimates of the joint MLE are very close to the true values. In the second scenario, we assume that

$$\sigma_\delta^2 = 1,\ \ \rho = 0.7,\ \ \sigma_{\epsilon_l}^2 = 0.1,\ \ \text{for}\ \ l = 0, \cdots, 5.$$

Again, the two-stage estimation gives very poor estimates, too small for σ_δ^2 and negative for ρ, while the joint MLE yields good estimation results. In the third scenario, we assume that

$$\sigma_\delta^2 = 0,\ \ \rho = 0.7,\ \ (\sigma_{\epsilon_0}^2, \sigma_{\epsilon_1}^2, \sigma_{\epsilon_2}^2, \sigma_{\epsilon_3}^2, \sigma_{\epsilon_4}^2, \sigma_{\epsilon_5}^2)' = (0.40, 0.35, 0.30, 0.25, 0.20, 0.15)'.$$

As shown in Table 3.1, the estimate for σ_δ^2 is close to zero for both methods. Because $\sigma_\delta^2 = 0$, two parameters, σ_δ^2 and ρ, are not separately identified, and thus the result for ρ is not meaningful. The estimates for $\sigma_{\epsilon_l}^2$ are quite

Table 3.1: Estimation results for the variance components of two methods based on simulated data.

		True Value	Joint MLE	Two-Stage
Scenario 1	σ^2_δ	1.00	1.12	0.19
	ρ	0.70	0.75	−0.26
	$\sigma^2_{\epsilon_0}$	0.40	0.33	0.61
	$\sigma^2_{\epsilon_1}$	0.35	0.39	0.52
	$\sigma^2_{\epsilon_2}$	0.30	0.26	0.36
	$\sigma^2_{\epsilon_3}$	0.25	0.23	0.32
	$\sigma^2_{\epsilon_4}$	0.20	0.23	0.38
	$\sigma^2_{\epsilon_5}$	0.15	0.11	0.54
Scenario 2	σ^2_δ	1.00	1.13	0.19
	ρ	0.70	0.76	−0.33
	$\sigma^2_{\epsilon_l}$	0.10	0.10	0.30
Scenario 3	σ^2_δ	0.00	0.00	0.00
	$\sigma^2_{\epsilon_0}$	0.40	0.31	0.32
	$\sigma^2_{\epsilon_1}$	0.35	0.37	0.35
	$\sigma^2_{\epsilon_2}$	0.30	0.25	0.26
	$\sigma^2_{\epsilon_3}$	0.25	0.22	0.22
	$\sigma^2_{\epsilon_4}$	0.20	0.20	0.19
	$\sigma^2_{\epsilon_5}$	0.15	0.15	0.15
Scenario 4	σ^2_δ	0.00	0.00	0.00
	$\sigma^2_{\epsilon_l}$	0.10	0.10	0.09

accurate with both methods. Finally, in the fourth scenario, we assume that

$$\sigma^2_\delta = 0, \;\; \rho = 0.7, \;\; \sigma^2_{\epsilon_l} = 0.1, \;\; \text{for } \; l = 0, \cdots, 5.$$

Again, the estimate for σ^2_δ is close to zero for both methods. The result for ρ is not meaningful due to an identifiability problem. The estimates for $\sigma^2_{\epsilon_l}$ are reasonably good with both methods.

Using simulated data, we have considered four different scenarios regarding the dependence structure of the errors. The joint MLE leads to robust estimation results for the parameters in the variance-covariance matrix. However, the two-stage estimation gives very poor estimates when the random effects are present.

Therefore, for more extensive simulation studies later in Section 3.4, we employ the joint MLE for estimating the parameters in Model (3.1), whose estimation results will be compared with those of an existing method without accounting for correlation.

3.3.4 Studies with Real Data

We apply the two methods to three different lysate arrays in order to see whether we have similar findings to the results with simulated data.

The first array, the lysate array of the protein $pThr308AKT$, is mentioned in Section 3.4.1. In this array, 96 ($s = 96$) samples are 2-fold serially diluted six times ($t = 6$) and spotted on a slide in triplicate ($r = 3$). We use the heteroscedastic error assumption and for the joint MLE, we combine $h = 12$ samples when estimating x_i, V_δ and V_ϵ. The estimation results for σ_δ^2, ρ and $\sigma_{\epsilon_l}^2$ from the two methods are given in Table 3.2. The results of two methods do not appear considerably different. The estimate for σ_δ^2 is very small for both methods, even considering the small scale of the measurements on this array. More precise values of the estimates are 2.60e-04 for the two-stage estimation and 3.95e-06 for the joint MLE. It seems that the random effects for the dilution series are not strong in this array. Then the estimate for ρ is not informative. Two methods give very similar estimates for $\sigma_{\epsilon_l}^2$, which supports our conclusion with simulated data that two methods give similar results when the random effects are negligible.

However, the arrays of AKT and $pmTor$ tell a different story. These arrays have been used in Section 3.2.1 and in both arrays, 40 samples ($s = 40$) are diluted eight times ($t = 8$) by a factor of two and spotted on a slide in duplicate ($r = 2$). For the Joint MLE, we combine $h = 10$ samples when estimating x_i, V_δ and V_ϵ. As shown in Table 3.2, for these two arrays, the random effects are not negligible and two methods lead to very different estimation results. The estimates for σ_δ^2 of the two-stage estimation are much smaller than the estimates of the joint MLE. For instance, in the array of AKT, the estimate from the two-stage estimation is $28 \cdot 10^3$, whereas the estimate of the joint MLE is $1675 \cdot 10^3$. In addition, the estimates for ρ from the two-stage estimation are negative, while the estimates from the joint MLE are large for both arrays. For example, in the array of AKT, the estimate from the two-stage estimation is -0.24, while the estimate from the joint MLE is 0.87. The results support our conclusion with simulated data that when the random effects exist, the two-stage estimation tends to underestimate σ_δ^2 and ρ. The scales for three arrays are very different due to the use of different spot-quantification softwares.

The application to real data shows similar patterns to those obtained with simulated data, and indicates that some lysate array data may have strong random effects of the dilution series. It seems quite necessary, therefore, to take into account the *non-i.i.d.* errors when analyzing the lysate array data.

Table 3.2: Estimation results for the variance components of two methods based on three lysate arrays.

		Joint MLE	Two-Stage
Array for $pThr308AKT$	σ^2_δ	0.00	0.00
	ρ	0.45	0.67
	$\sigma^2_{\epsilon_0}$	0.47	0.49
	$\sigma^2_{\epsilon_1}$	0.13	0.14
	$\sigma^2_{\epsilon_2}$	0.01	0.01
	$\sigma^2_{\epsilon_3}$	0.07	0.06
	$\sigma^2_{\epsilon_4}$	0.15	0.13
	$\sigma^2_{\epsilon_5}$	0.19	0.18
Array for AKT	σ^2_δ	$1675 \cdot 10^3$	$28 \cdot 10^3$
	ρ	0.87	-0.24
	$\sigma^2_{\epsilon_0}$	$7 \cdot 10^6$	$1 \cdot 10^6$
	$\sigma^2_{\epsilon_1}$	$23 \cdot 10^6$	$3 \cdot 10^6$
	$\sigma^2_{\epsilon_2}$	$13 \cdot 10^6$	$2 \cdot 10^6$
	$\sigma^2_{\epsilon_3}$	$8 \cdot 10^6$	$2 \cdot 10^6$
	$\sigma^2_{\epsilon_4}$	$3 \cdot 10^6$	$3 \cdot 10^6$
	$\sigma^2_{\epsilon_5}$	$0.5 \cdot 10^6$	$1 \cdot 10^6$
	$\sigma^2_{\epsilon_6}$	$0.1 \cdot 10^6$	$0.3 \cdot 10^6$
	$\sigma^2_{\epsilon_7}$	$0.3 \cdot 10^6$	$0.2 \cdot 10^6$
Array for $pmTor$	σ^2_δ	$305 \cdot 10^3$	$0.22 \cdot 10^3$
	ρ	0.75	-0.44
	$\sigma^2_{\epsilon_0}$	$5 \cdot 10^6$	$3 \cdot 10^6$
	$\sigma^2_{\epsilon_1}$	$3 \cdot 10^6$	$1 \cdot 10^6$
	$\sigma^2_{\epsilon_2}$	$0.3 \cdot 10^6$	$0.4 \cdot 10^6$
	$\sigma^2_{\epsilon_3}$	$0.5 \cdot 10^6$	$1 \cdot 10^6$
	$\sigma^2_{\epsilon_4}$	$0.8 \cdot 10^6$	$0.9 \cdot 10^6$
	$\sigma^2_{\epsilon_5}$	$0.5 \cdot 10^6$	$1 \cdot 10^6$
	$\sigma^2_{\epsilon_6}$	$0.3 \cdot 10^6$	$0.7 \cdot 10^6$
	$\sigma^2_{\epsilon_7}$	$0.1 \cdot 10^6$	$0.5 \cdot 10^6$

3.3.5 An Example with Joint MLE

The Joint MLE mainly consists of two parts: a one-step iteration procedure and a refining procedure to attain better estimates. In this section, using a simulated data set we will show that the one-step iteration (or even several iterations) sometimes produces poor estimates, leaving room for improvement with a refining procedure.

We generate data from Model (3.1) using the following values:

$$s = 96, r = 3, t = 6, \underline{\beta} = (1, 1.5, 1)', \sigma_\delta^2 = 0.1^2, \rho = 0.7,$$
$$(\sigma_{\epsilon_0}^2, \sigma_{\epsilon_1}^2, \cdots, \sigma_{\epsilon_5}^2)' = (0.1^2, 0.1^2, \cdots, 0.1^2)'.$$

The concentration levels are generated from the skew-normal distribution with location, scale, and shape parameters (3,2,-3). We combine $h = 24$ samples when estimating x_i, V_δ and V_ϵ.

Figure 3.2 shows the estimation results after several iterations. The first plot compares the relative concentration estimates, $\hat{x}_i - \hat{x}_{ref}$, with the true levels. The results after one, three, seven, and ten iterations are presented. We see that the estimates are more deviated from the true levels after ten iterations than after one iteration. The solid line is the 45-degree line. The plot in the upper right corner shows the trend of D_{tot} with the number of iterations. As in the previous chapter, D_{tot} is defined as the sum of the absolute differences between the estimated relative concentration levels and the true relative levels. Overall, the value of D_{tot} increases with the number of iterations. The two plots in the bottom row show the trend of σ_δ^2 estimates and ρ estimates, with the number of iterations, respectively. The dashed lines indicate the true values in both plots. The estimates are quite far off from the true values after one-step iteration or after several iterations.

This example shows that repeated iterations between Steps 1 and 2 in the joint MLE procedure do not necessarily lead to good results. That is why we use Step 4 as a refining procedure.

3.4 Simulation Studies

Now we carry out several simulation studies to evaluate the overall performance of the joint MLE (not only the estimation of the variance components but also the estimation of x_i and $\underline{\beta}$), in comparison with the single-step least squares procedure on all the samples used in [10]. We consider three different cases of error structures and curve parameters.

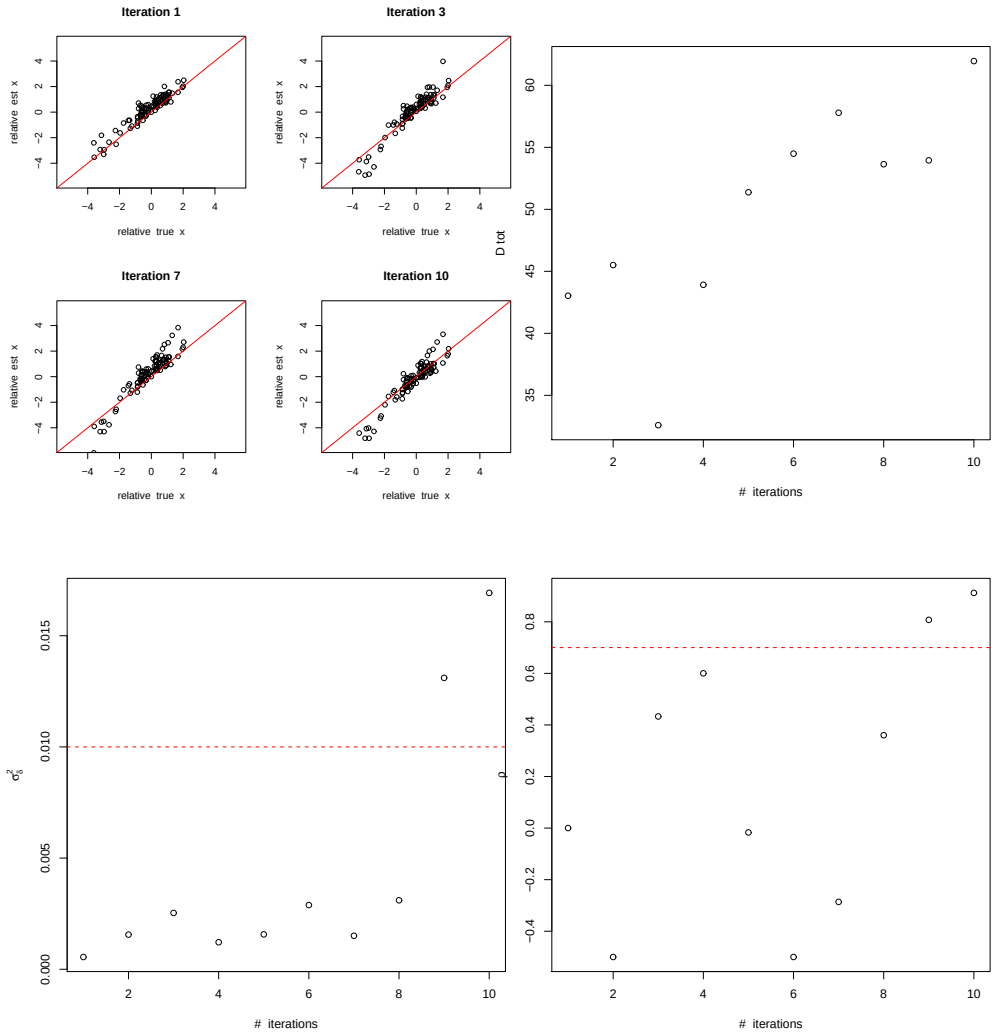

Figure 3.2: The upper left corner corresponds to the comparison of the relative concentration estimates with the true relative levels. The upper right corner corresponds to D_{tot} values along with the number of iterations. Two plots in the bottom row correspond to σ_δ^2 estimates and ρ estimates, with the number of iterations, respectively, and the dashed lines indicate the true values in both plots.

3.4.1 Implementation Details and Assessment Criteria

When generating simulation data sets, we try to imitate real data, and thus we use as references the results of the real data analysis that has been performed in the previous chapter and in Section 3.3.4. Throughout the studies, we assume that the number of samples is 96 ($s = 96$), the number of replicates is 3 ($r = 3$), and the number of dilution levels is either 6 or 8 ($t = 6$ or 8). We generate x_i from the skew-normal distribution with location, scale, and shape parameters (3,2,-3). We use the identical set of x_i ($i = 1, \cdots, s$) for 100 simulation data sets throughout the studies. For the joint MLE, we combine $h = 12$ samples when estimating x_i, V_δ and V_ϵ.

For each simulation data set, we compute the same performance measure, D_{tot}, as in the previous chapter. The middle sample, in terms of the true concentration level, is used as the reference again. The definition of D_{tot} is as follows:

$$D_{tot} = \sum_{i=1}^{s} |\tilde{D}_i - D_i|,$$

where $\tilde{D}_i = \tilde{x}_i - \tilde{x}_{ref}$, $D_i = x_i - x_{ref}$, $\tilde{x}_i$ and $\tilde{x}_{ref}$ are the estimates for the concentration levels of the i^{th} sample and of the reference sample, respectively, and x_i and x_{ref} are their true values. The smaller D_{tot} is, the more desirable the result is.

In addition, the same Winsorization rule as in Section 2.4.1 is applied.

3.4.2 Simulation Case 1

We consider a case where the random effects of the dilution series are strong and the errors are heteroscedastic. We assume that

$$\sigma_\delta^2 = 1, \;\; \rho = 0.7, \;\; (\sigma_{\epsilon_0}^2, \sigma_{\epsilon_1}^2, \sigma_{\epsilon_2}^2, \sigma_{\epsilon_3}^2, \sigma_{\epsilon_4}^2, \sigma_{\epsilon_5}^2)' = (0.40, 0.35, 0.30, 0.25, 0.20, 0.15)'.$$

In addition, we use $\underline{\beta} = (10, 15, 1)'$. Figure 3.3 shows the distributions of D_{tot}, based on 100 simulation data sets. The first plot corresponds to the joint MLE and the second plot corresponds to the single-step procedure used in [10]. Evidently, in this case the joint MLE leads to smaller D_{tot} than the single-step procedure. It indicates that when the random effects exist, the method that ignores these effects may yield poor estimation results for x_i. The summary statistics of D_{tot} for two methods, are given in Table 3.3.

Table 3.3 presents the estimation results for the curve parameters, including the summary statistics and the mean squared error based on 100

estimates for each of the curve parameters. For each of the parameters, β_1, β_2, and β_3, the joint MLE produces much smaller mean squared error than the single-step procedure.

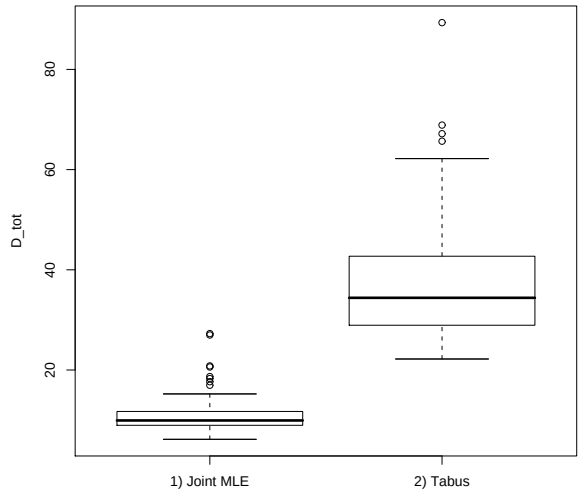

Figure 3.3: Box-and-Whisker plots of the distribution of D_{tot} in simulation Case 1.

For further exploration, we examine two individual trials. Two plots in the upper panel of Figure 3.4 are based on the data set, for which the joint MLE shows the largest D_{tot} value, 27.27 (D_{tot} value of the single-step procedure in this trial is 32.33). Two plots in the lower panel correspond to the data set, for which the single-step procedure shows the largest D_{tot} value, 89.34 (D_{tot} value of the joint MLE in this trial is 8.45). The plots in the first column present the true curve and the data points, and the plots in the second column compare two fitted curves of two methods with the true curve for each trial. The fitted curves have been shifted horizontally so that the x(-axis) value at the inflection point is zero, which is reasonable because we are more interested in the relative value of the concentration levels. In both trials, the fitted curve of the single-step procedure is slightly off the true curve, while the joint MLE produces a good curve fit even for the worst data set. It seems that the single-step procedure is more likely to fail to accurately estimate the curve in the tails than the joint MLE, which will surely affect the estimation for x_i.

With the joint MLE, we can estimate the variance components. The estimation results based on 100 data sets are given in Table 3.4. When

Table 3.3: Summary statistics of D_{tot} and the curve parameter estimates in simulation Case 1.

		True Value	Joint MLE	Tabus
D_{tot}	Min.		6.17	22.20
	Median		9.30	34.41
	Mean		10.91	38.09
	Max.		27.27	89.34
β_1	Min.	10.00	9.80	9.37
	Median		9.99	9.70
	Mean		10.01	9.72
	Max.		10.37	10.14
	MSE		0.01	0.10
β_2	Min.	15.00	14.60	15.02
	Median		14.87	15.95
	Mean		14.88	15.94
	Max.		15.27	16.73
	MSE		0.04	1.02
β_3	Min.	1.00	0.96	0.83
	Median		1.00	0.90
	Mean		0.99	0.90
	Max.		1.03	0.99
	MSE		0.00	0.01

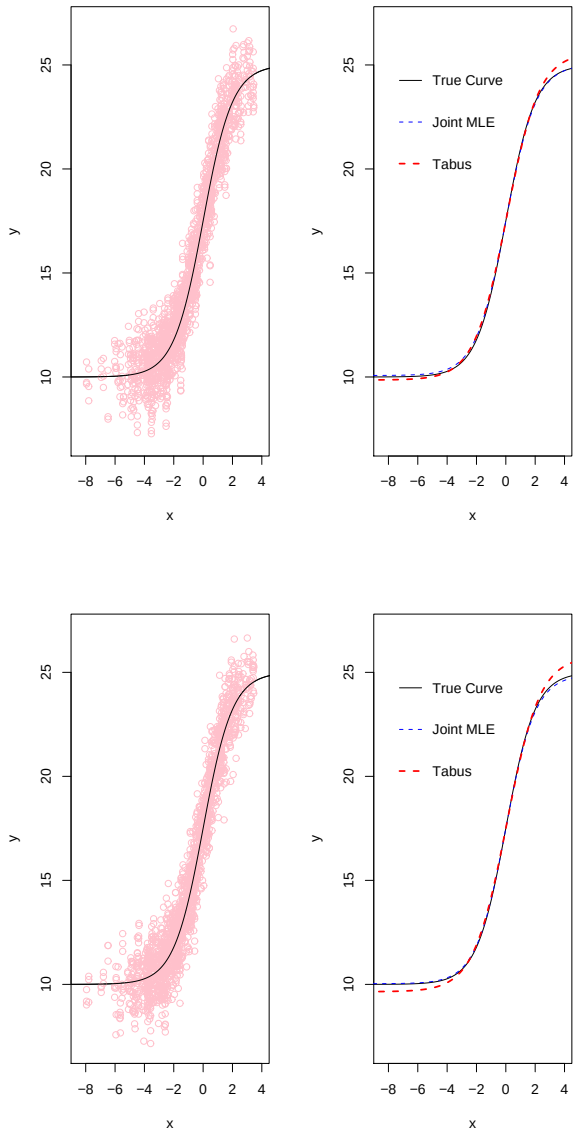

Figure 3.4: True curve versus fitted curves of two methods in simulation Case 1: two plots in the upper panel correspond to the worst trial for the joint MLE, and two plots in the lower panel correspond to the worst trial for the single-step procedure. The plots in the first column present the true curve and the data points, and the plots in the second column compare two fitted curves of two methods with the true curve.

compared to the true values, the joint MLE works well for estimating the variance components in Model (3.1).

In simulation Case 1, the random effects of the dilution series are strong. We have found that the joint MLE leads to much more accurate estimates for both x_i and $\underline{\beta}$ than the single-step procedure used in [10]. It indicates that when the random effects are present, the method that ignores these effects can yield poor estimation results.

3.4.3 Simulation Case 2

We consider a case where the random effects of the dilution series are not present and the errors are homoscedastic, that is, the errors are *i.i.d.* It turns out that the proposed joint MLE gives comparable estimation results to the single-step procedure, which is designed under the *i.i.d.* model. We assume that

$$\sigma_\delta^2 = 0, \;\; \rho = 0.7, \;\; \sigma_{\epsilon_l}^2 = 50, \;\; l = 0, \cdots, 5.$$

In addition, we assume that $\underline{\beta} = (100, 250, 0.7)'$. The choice for the values of the curve parameters takes the cue from the real data analysis performed in Section 3.3.4. Particularly, in *pmTor* array, we have noticed that the median of $\hat{\sigma}_{\epsilon_l}$ is roughly 5% of $\hat{\beta}_2 - \hat{\beta}_1$.

The distributions of D_{tot}, based on 100 data sets, are given in Figure 3.5. Two methods produce very similar results regarding D_{tot}. However, as shown in Table 3.5, the joint MLE gives much smaller mean squared error for the curve parameters than the single-step procedure.

The variance component estimates of the joint MLE, based on 100 trials, are given in Table 3.6. In terms of the medians or means, the estimates for σ_δ^2 and $\sigma_{\epsilon_l}^2$ are reasonably accurate. Due to an identifiability problem, the estimate for ρ is not meaningful in this case.

In simulation Case 2, we assume that the measurement errors are *i.i.d.* The joint MLE and the single-step procedure give very similar estimation results for x_i. If we take it into account that the single-step procedure of [10] is designed for the *i.i.d.* case, the joint MLE has a robust performance.

3.4.4 Simulation Case 3

To mimic the data we saw from the protein, *pmTor* analyzed in Section 3.3.4, we use the following values:

$$\underline{\beta} = (11187, 26007, 0.7)', \;\; \sigma_\delta^2 = 305054, \;\; \rho = 0.75,$$

Table 3.4: Estimation results for the variance components in simulation Case 1.

		True Value	Joint MLE
σ^2_δ	Min.	1.00	0.66
	Median		0.96
	Mean		0.96
	Max.		1.39
ρ	Min.	0.70	0.51
	Median		0.67
	Mean		0.67
	Max.		0.81
$\sigma^2_{\epsilon_0}$	Min.	0.40	0.26
	Median		0.37
	Mean		0.37
	Max.		0.50
$\sigma^2_{\epsilon_1}$	Min.	0.35	0.26
	Median		0.33
	Mean		0.33
	Max.		0.44
$\sigma^2_{\epsilon_2}$	Min.	0.30	0.23
	Median		0.29
	Mean		0.29
	Max.		0.37
$\sigma^2_{\epsilon_3}$	Min.	0.25	0.18
	Median		0.24
	Mean		0.24
	Max.		0.30
$\sigma^2_{\epsilon_4}$	Min.	0.20	0.15
	Median		0.20
	Mean		0.20
	Max.		0.25
$\sigma^2_{\epsilon_5}$	Min.	0.15	0.09
	Median		0.13
	Mean		0.13
	Max.		0.18

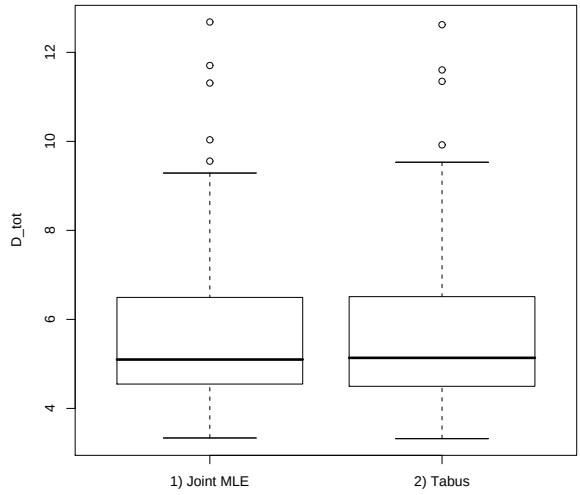

Figure 3.5: Box-and-Whisker plots of the distribution of D_{tot} in simulation Case 2.

Table 3.5: Summary statistics of D_{tot} and the curve parameter estimates in simulation Case 2.

		True Value	Joint MLE	Tabus
D_{tot}	Min.		3.34	3.32
	Median		5.10	5.14
	Mean		5.74	5.70
	Max.		12.68	12.62
β_1	Min.	100.00	98.88	68.73
	Median		100.50	99.97
	Mean		100.50	99.40
	Max.		102.80	102.00
	MSE		0.86	18.40
β_2	Min.	250.00	241.60	245.20
	Median		247.90	250.50
	Mean		247.60	253.50
	Max.		253.60	416.80
	MSE		11.98	533.76
β_3	Min.	0.70	0.68	0.40
	Median		0.70	0.70
	Mean		0.70	0.69
	Max.		0.73	0.72
	MSE		0.00	0.00

Table 3.6: Estimation results for the variance components in simulation Case 2.

		True Value	Joint MLE
σ^2_δ	Min.	0.00	0.00
	Median		0.27
	Mean		0.42
	Max.		2.54
$\sigma^2_{\epsilon_l}$	Min.	50.00	42.49
	Median		46.98
	Mean		47.34
	Max.		78.96

$$(\sigma^2_{\epsilon_0}, \sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}, \sigma^2_{\epsilon_3}, \sigma^2_{\epsilon_4}, \sigma^2_{\epsilon_5}, \sigma^2_{\epsilon_6}, \sigma^2_{\epsilon_7})'$$

$$= (4542884, 2883969, 342657, 471029, 818083, 506195, 321632, 49902)'$$

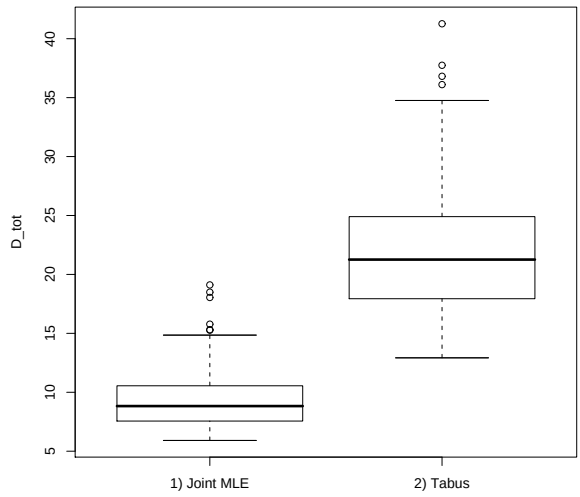

Figure 3.6: Box-and-Whisker plots of the distribution of D_{tot} in simulation Case 3.

Figure 3.6 shows the distributions of D_{tot}, based on 100 data sets. Overall, the joint MLE leads to smaller D_{tot} than the single-step procedure. Again, the results imply that we should use the method that incorporates the random effects when the effects exist. Besides, Table 3.7 shows that the joint MLE leads to more accurate curve parameter estimation as well.

We explore an individual trial in which both the joint MLE and the single-

Table 3.7: Summary statistics of D_{tot} and the curve parameter estimates in simulation Case 3.

		True Value	Joint MLE	Tabus
D_{tot}	Min.		5.91	12.92
	Median		8.83	21.26
	Mean		9.52	22.26
	Max.		19.11	41.27
β_1	Min.	$112 \cdot 10^2$	$110 \cdot 10^2$	$109 \cdot 10^2$
	Median		$112 \cdot 10^2$	$112 \cdot 10^2$
	Mean		$112 \cdot 10^2$	$112 \cdot 10^2$
	Max.		$113 \cdot 10^2$	$114 \cdot 10^2$
	MSE		$36 \cdot 10^2$	$89 \cdot 10^2$
β_2	Min.	$260 \cdot 10^2$	$247 \cdot 10^2$	$254 \cdot 10^2$
	Median		$257 \cdot 10^2$	$263 \cdot 10^2$
	Mean		$257 \cdot 10^2$	$263 \cdot 10^2$
	Max.		$266 \cdot 10^2$	$277 \cdot 10^2$
	MSE		$2341 \cdot 10^2$	$3017 \cdot 10^2$
β_3	Min.	0.70	0.67	0.66
	Median		0.70	0.69
	Mean		0.70	0.69
	Max.		0.73	0.72
	MSE		0.00	0.00

step procedure show the worst performance in terms of D_{tot}: the joint MLE gives 19.11 and the single-step procedure gives 41.27. Figure 3.7 presents the true curve and the fitted curves of two methods for this worst trial. The single-step procedure fails to accurately estimate the upper part of the curve. On the other hand, the joint MLE gives a reasonably good curve fit even in this trial.

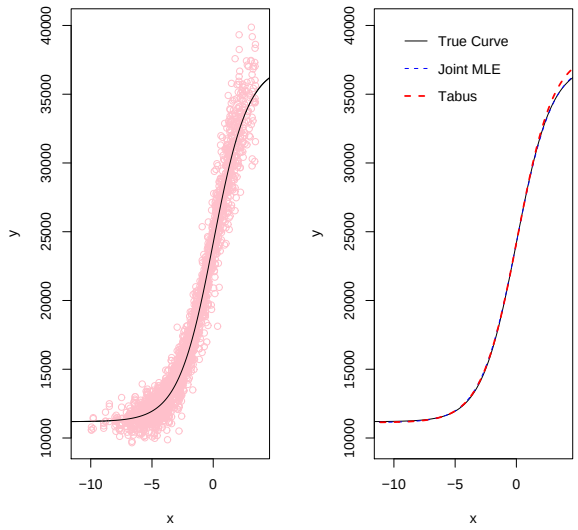

Figure 3.7: True curve versus fitted curves of two methods in simulation Case 3: these plots are based on the worst simulation trial for both methods in terms of D_{tot}. The left plot presents the true curve and the data points, and the right plot compares the fitted curves of two methods with the true curve.

The variance component estimates for the joint MLE are given in Table 3.8. In terms of the medians or means, the estimates appear to be reasonably acceptable, although they are not as precise as in the previous two cases. The variation of the 100 estimates is quite large.

In simulation Case 3, we assume that the random effects of the dilution series are strong and the error variances are large. The joint MLE still leads to more accurate estimates (both for x_i and $\underline{\beta}$) than the single-step procedure.

Table 3.8: Estimation results for the variance components in simulation Case 3.

		True Value	Joint MLE
σ^2_δ	Min.	$0.3 \cdot 10^6$	$0.1 \cdot 10^6$
	Median		$0.3 \cdot 10^6$
	Mean		$0.3 \cdot 10^6$
	Max.		$0.6 \cdot 10^6$
ρ	Min.	0.75	-0.26
	Median		0.65
	Mean		0.58
	Max.		0.92
$\sigma^2_{\epsilon_0}$	Min.	$5 \cdot 10^6$	$2 \cdot 10^6$
	Median		$4 \cdot 10^6$
	Mean		$4 \cdot 10^6$
	Max.		$5 \cdot 10^6$
$\sigma^2_{\epsilon_1}$	Min.	$3 \cdot 10^6$	$2 \cdot 10^6$
	Median		$3 \cdot 10^6$
	Mean		$3 \cdot 10^6$
	Max.		$4 \cdot 10^6$
$\sigma^2_{\epsilon_2}$	Min.	$0.3 \cdot 10^6$	$0.2 \cdot 10^6$
	Median		$0.3 \cdot 10^6$
	Mean		$0.3 \cdot 10^6$
	Max.		$0.5 \cdot 10^6$
$\sigma^2_{\epsilon_3}$	Min.	$0.5 \cdot 10^6$	$0.3 \cdot 10^6$
	Median		$0.5 \cdot 10^6$
	Mean		$0.5 \cdot 10^6$
	Max.		$0.7 \cdot 10^6$
$\sigma^2_{\epsilon_4}$	Min.	$0.8 \cdot 10^6$	$0.6 \cdot 10^6$
	Median		$0.8 \cdot 10^6$
	Mean		$0.8 \cdot 10^6$
	Max.		$1.4 \cdot 10^6$
$\sigma^2_{\epsilon_5}$	Min.	$0.5 \cdot 10^6$	$0.4 \cdot 10^6$
	Median		$0.5 \cdot 10^6$
	Mean		$0.5 \cdot 10^6$
	Max.		$0.8 \cdot 10^6$
$\sigma^2_{\epsilon_6}$	Min.	$0.3 \cdot 10^6$	$0.2 \cdot 10^6$
	Median		$0.3 \cdot 10^6$
	Mean		$0.3 \cdot 10^6$
	Max.		$0.6 \cdot 10^6$
$\sigma^2_{\epsilon_7}$	Min.	$0.1 \cdot 10^6$	$0.0 \cdot 10^6$
	Median		$0.1 \cdot 10^6$
	Mean		$0.1 \cdot 10^6$
	Max.		$0.3 \cdot 10^6$

3.4.5 Confidence Intervals

Finally, we compute the confidence intervals for the relative concentration levels, and investigate coverage probability performances of the joint MLE and the single-step procedure. When the errors are *non-i.i.d.*, the ordinary least squares estimation used for the single-step procedure is well known to produce a biased variance estimate, resulting in incorrect confidence intervals.

In the previous chapter, we have shown that the convergence rate of the curve parameter estimate is faster than that of the concentration estimate: the convergence rate of the former is $O_p\big((srt)^{-1/2}\big)$, while the convergence rate of the latter is $O_p\big((rt)^{-1/2}\big)$. We can expect that the convergence rate of the variance-covariance matrix, $cov(\underline{\Upsilon}) = I_s \otimes \big(V_\delta \otimes J_t + I_r \otimes V_\epsilon\big)$, in Model (3.1) will be also faster than that of the concentration level. Therefore, the estimate of x_i for the joint MLE will be as efficient as in a problem with known curve parameters and known variance-covariance matrix. Then the variance of $\tilde{x}_i$ takes the following form:

$$var(\tilde{x}_i) = \Big[\big(\frac{\partial g_i}{\partial x_i}\big)' \, Q^{-1} \, \big(\frac{\partial g_i}{\partial x_i}\big) \Big]^{-1},$$

where g_i is the rt by 1 vector for the mean effects of the i^{th} sample and $Q = V_\delta \otimes J_t + I_r \otimes V_\epsilon$. Using the joint MLE, we can construct a $100(1-\alpha)\%$ asymptotic confidence interval for $x_i - x_{ref}$ as follows:

$$CI_1 = (\tilde{x}_i - \tilde{x}_{ref}) \pm z_{\alpha/2} \, \big(\widehat{var}(\tilde{x}_i) + \widehat{var}(\tilde{x}_{ref})\big)^{1/2},$$

where $\widehat{var}(\tilde{x}_i)$ and $\widehat{var}(\tilde{x}_{ref})$ are the estimates for $var(\tilde{x}_i)$ and $var(\tilde{x}_{ref})$, which are obtained by replacing the true values with the corresponding estimates of the joint MLE.

[10] did not provide a variance estimate, but we can use the asymptotic result of Theorem 1 in the previous chapter to compute the variance, because we expect Theorem 1 to hold for the single-step estimates. Then the variance of $\hat{x}_i$ is given by

$$var(\hat{x}_i) = \frac{\sigma^2}{r} \Bigg[\sum_{l=0}^{t-1} \frac{{\beta_2}^2 {\beta_3}^2 e^{-2\beta_3(x_i - l)}}{\big(1 + e^{-\beta_3(x_i - l)}\big)^4} \Bigg]^{-1},$$

which takes the same form as

$$var(\hat{x}_i) = \sigma^2 \Big[\big(\frac{\partial g_i}{\partial x_i}\big)' \, \big(\frac{\partial g_i}{\partial x_i}\big) \Big]^{-1}.$$

Using the single-step procedure, we can construct a $100(1-\alpha)\%$ asymptotic confidence interval for $x_i - x_{ref}$ as follows:

$$CI_2 = (\hat{x}_i - \hat{x}_{ref}) \pm z_{\alpha/2} \left(\widehat{var}(\hat{x}_i) + \widehat{var}(\hat{x}_{ref})\right)^{1/2},$$

where $\widehat{var}(\hat{x}_i)$ and $\widehat{var}(\hat{x}_{ref})$ are the estimates for $var(\hat{x}_i)$ and $var(\hat{x}_{ref})$, which are obtained by replacing the true values with the corresponding estimates of the single-step procedure. We estimate σ^2 by

$$\hat{\sigma}^2 = \sum_{ijl} \left(y_{ijl} - g(\underline{\hat{\beta}}, \hat{x}_i - l)\right) \,/\, (srt - s - 3),$$

where $g(\underline{\beta}, x_i - l) = \beta_1 + \frac{\beta_2}{1+e^{-\beta_3(x_i-l)}}$.

For the simulation setting of Case 1, we compute 90% confidence intervals. We use 100 data sets and for each data set we compute 95 confidence intervals on pairwise difference ($x_i - x_{ref},\ i = 1, 2, ..., 96,\ i \neq ref$). Again, as a reference, we use the sample whose true concentration level is the median of 96 x_i's. The coverage probability is defined as the proportion of the intervals that contain the true value among 100×95 intervals. In addition, we define the interval length for each of 95 confidence intervals on pairwise difference, as the average length of the corresponding 100 confidence intervals, which makes sense because we use the identical x_i's for all 100 data sets. It turns out that the single-step procedure has a very low coverage probability, 0.32, due to the severe bias in the estimation for $x_i - x_{ref}$, as already shown in Figure 3.3, and also generally short intervals, as shown in Figure 3.8. On the other hand, the coverage probability for the joint MLE is 0.90, which is equal to the nominal level. Overall, the joint MLE results in wider confidence intervals than the single-step procedure as shown in Figure 3.8. The average interval length of all 100×95 intervals is 0.44 for the joint MLE, and 0.39 for the single-step procedure. Three samples whose x_i estimates are in the tails have been excluded from Figure 3.8 for both methods.

Next, we use the simulation setting of Case 2, and compute 90% confidence intervals. In this case, the measurement errors are *i.i.d.* The coverage probabilities of the joint MLE and the single-step procedure are 0.89 and 0.90, respectively. In the *i.i.d.* case, both methods have the coverage probabilities that are very close to the nominal level. However, as shown in Figure 3.9, the joint MLE produces slightly shorter confidence intervals than the single-step procedure. Again, the three samples that have x_i estimates in the tails are not included in Figure 3.9. The average interval lengths are

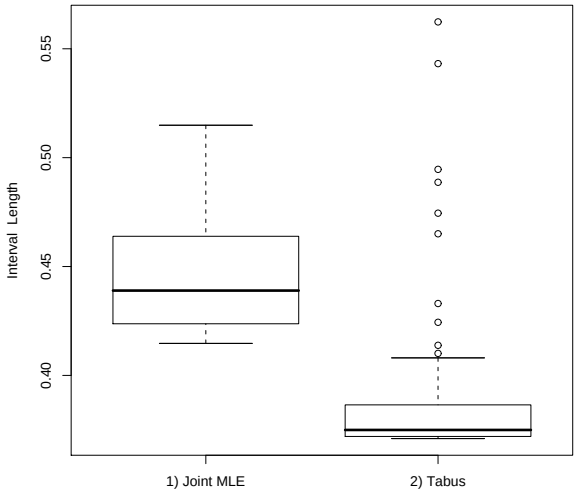

Figure 3.8: Confidence interval lengths for $x_i - x_{ref}$ in simulation Case 1.

Table 3.9: Coverage probabilities (the nominal level is 0.90) and average confidence interval lengths.

		Joint MLE	Tabus
simulation Case 1	Coverage Prob.	0.90	0.32
	Average Interval Length	0.44	0.39
	Median Interval Length	0.44	0.37
simulation Case 2	Coverage Prob.	0.89	0.90
	Average Interval Length	0.23	0.24
	Median Interval Length	0.23	0.24
simulation Case 3	Coverage Prob.	0.88	0.56
	Average Interval Length	0.36	0.38
	Median Interval Length	0.34	0.36

0.23 for the joint MLE, and 0.24 for the single-step procedure, as presented in Table 3.9.

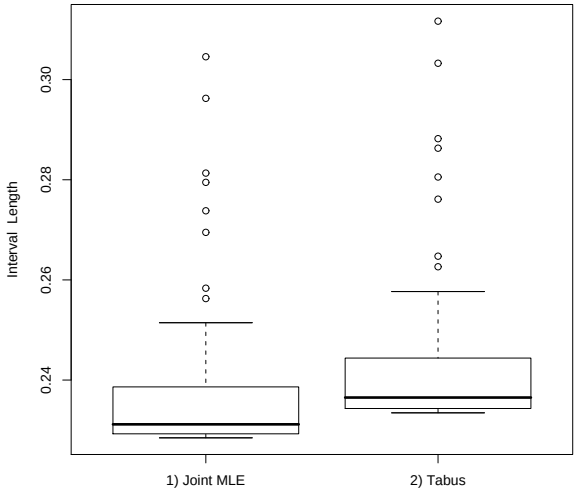

Figure 3.9: Confidence interval lengths for $x_i - x_{ref}$ in simulation Case 2.

Based on the simulation setting of Case 3, we compute 90% confidence intervals. The coverage probabilities of the joint MLE and the single-step procedure are 0.88 and 0.56, respectively. For the joint MLE, the coverage probability is slightly below the nominal level, but still considered acceptable. For the single-step procedure, the coverage probability is far below the nominal level. Figure 3.10 shows the confidence interval lengths of two methods. The three samples whose x_i estimates are in the tails are excluded from this figure. The average lengths are 0.36 for the joint MLE, and 0.38 for the single-step procedure.

In this section, we investigate the joint MLE and the single-step procedure in terms of the coverage probability and the length of confidence intervals in three different simulation settings. The joint MLE has the coverage probability that is very close to the nominal level in all three cases. However, the single-step procedure has a very low coverage probability in the cases with *non-i.i.d.* errors and strong random effects of the dilution series.

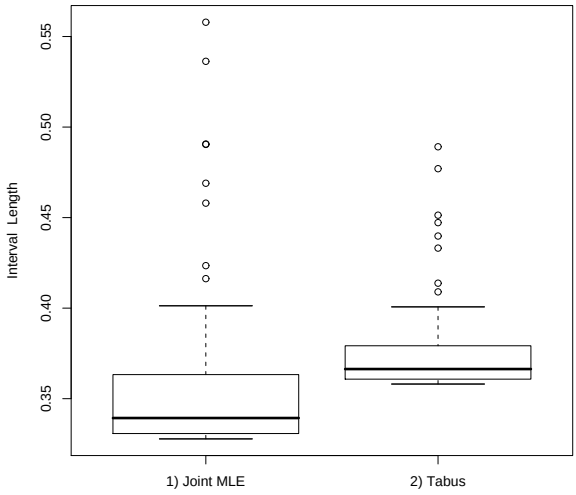

Figure 3.10: Confidence interval lengths for $x_i - x_{ref}$ in simulation Case 3.

3.4.6 Discussion and Future Work

A model that takes into account possible *non-i.i.d.* error structures appears very necessary for analyzing the lysate array data, because the examination of real data warns us that the *i.i.d.* assumption may be not reasonable in reality. In this chapter, we introduce a new nonlinear mixed effects model that allows for the dependence structure of the errors. Based on the new model, we propose a method to approximate the joint maximum likelihood estimator of all the parameters. Using simulation studies on various error structures, we show that the proposed method leads to more accurate estimates for both the protein concentration levels and the curve parameters than the single-step least squares procedure of [10], when strong random effects of the dilution series are present. More importantly, the joint MLE gives much better confidence intervals under various simulation settings, whereas the procedure based on the *i.i.d.* error assumption often leads to low coverage probabilities.

The dependence structure in our new model makes the likelihood function more complicated. Also the likelihood involves a high-dimensional parameter space due to the nature of the lysate array data. We considered a method to approximate the joint maximum likelihood estimator of all the parameters, including the variance components, and showed that the results are promising. As an alternative, in future work, we will employ two other

methods that may improve computational efficiency and stability. The first is the EM algorithm and the second is a Bayesian approach with MCMC.

In addition, based on the new model we will extend the multi-step procedure proposed in the previous chapter. We estimate the parameters on the group basis, and pool the group-based estimates for the curve parameters and variance components using an appropriate weight matrix. Based on these pooled estimates, we finally estimate the concentration levels. In doing this, we avoid working with parameters of increasing dimensions, which makes large sample inferences easy and helps the numerical computation stable.

APPENDIX A

We give the explicit expressions of B_{vw} and M_{vw}, $1 \leq v, w \leq (k+3)$, shown in Section 2.2.3 in Chapter 2. Let $\eta_i = e^{-\beta_3^q(x_i^q - l)}$, $\eta_u = e^{-\beta_3^q(x_u^q - l)}$ and $\sum_{i,l} = \sum_{i=1}^{k}\sum_{l=0}^{t-1}$. For $u = 1, 2, ..., k$, $a = 4, 5, ..., (k+2)$ and $b = (a+1), (a+2), ..., (k+3)$, we have

$$B_{11} = tk,\ B_{12} = \sum_{i,l} \frac{1}{1+\eta_i},\ B_{13} = \sum_{i,l} \frac{\beta_2^q(x_i^q - l)\eta_i}{\left(1+\eta_i\right)^2},\ B_{1(u+3)} = \sum_{l=0}^{t-1} \frac{\beta_2^q\beta_3^q\eta_u}{\left(1+\eta_u\right)^2},$$

$$B_{22} = \sum_{i,l} \frac{1}{\left(1+\eta_i\right)^2},\ B_{23} = \sum_{i,l} \frac{\beta_2^q(x_i^q - l)\eta_i}{\left(1+\eta_i\right)^3},\ B_{2(u+3)} = \sum_{l=0}^{t-1} \frac{\beta_2^q\beta_3^q\eta_u}{\left(1+\eta_u\right)^3},$$

$$B_{33} = \sum_{i,l} \frac{(\beta_2^q)^2(x_i^q - l)^2\eta_u^2}{\left(1+\eta_u\right)^4},\ B_{3(u+3)} = \sum_{l=0}^{t-1} \frac{(\beta_2^q)^2\beta_3^q(x_u^q - l)\eta_i^2}{\left(1+\eta_u\right)^4},$$

$$B_{(u+3)(u+3)} = \sum_{l=0}^{t-1} \frac{(\beta_2^q)^2(\beta_3^q)^2\eta_u^2}{\left(1+\eta_u\right)^4},\ B_{ab} = 0,\ B_{vw} = B_{wv}.$$

In addition to the common variance assumption of the errors, we also assume that the measurements are independent across different biological samples or different replicates, but dependent within the same replicate for the same biological sample having the correlation coefficient ρ, that is, $var(\epsilon_{ijl}) = \sigma^2$, $cov(\epsilon_{ijl}, \epsilon_{i^*jl}) = 0$, $cov(\epsilon_{ijl}, \epsilon_{ij^*l}) = 0$, and $cov(\epsilon_{ijl}, \epsilon_{ijl^*}) = \rho\sigma^2$, $i \neq i^*$, $j \neq j^*$,
$l \neq l^*$. Let $\eta_i^* = e^{-\beta_3^q(x_i^q - l^*)}$, $\sum_{i,l,l^*} = \sum_{i=1}^{k}\sum_{l=0}^{t-1}\sum_{l^*=l+1}^{t-1}$, and $\sum_{l,l^*} = \sum_{l=0}^{t-1}\sum_{l^*=l+1}^{t-1}$. Then, for $u = 1, 2, ..., k$, $a = 4, 5, ..., (k+2)$ and $b = (a+1), (a+2), ..., (k+3)$, we have

$$M_{11} = tk + t(t-1)k\rho,\ M_{12} = \sum_{i,l} \frac{1}{1+\eta_i} + (t-1)\rho\sum_{i,l} \frac{1}{1+\eta_i},$$

$$M_{13} = \sum_{i,l} \frac{\beta_2^q(x_i^q - l)\eta_i}{\left(1+\eta_i\right)^2} + (t-1)\rho\sum_{i,l} \frac{\beta_2^q(x_i^q - l)\eta_i}{\left(1+\eta_i\right)^2},$$

$$M_{1(u+3)} = \sum_{l=0}^{t-1} \frac{\beta_2^q \beta_3^q \eta_u}{\left(1+\eta_u\right)^2} + (t-1)\rho \sum_{l=0}^{t-1} \frac{\beta_2^q \beta_3^q \eta_u}{\left(1+\eta_u\right)^2},$$

$$M_{22} = \sum_{i,l} \frac{1}{\left(1+\eta_i\right)^2} + 2\ \rho \sum_{i,l,l^*} \frac{1}{\left(1+\eta_i\right)\left(1+\eta_i^*\right)},$$

$$M_{23} = \sum_{i,l} \frac{\beta_2^q (x_i^q - l)\eta_i}{\left(1+\eta_i\right)^3} + 2\ \rho \sum_{i,l,l^*} \frac{\beta_2^q (x_i^q - l)\eta_i}{\left(1+\eta_i\right)^2\left(1+\eta_i^*\right)},$$

$$M_{2(u+3)} = \sum_{l=0}^{t-1} \frac{\beta_2^q \beta_3^q \eta_u}{\left(1+\eta_u\right)^3} + 2\ \rho \sum_{l,l^*} \frac{\beta_2^q \beta_3^q \eta_u}{\left(1+\eta_u\right)^2\left(1+\eta_u^*\right)},$$

$$M_{33} = \sum_{i,l} \frac{(\beta_2^q)^2 (x_i^q - l)^2 \eta_i^2}{\left(1+\eta_i\right)^4} + 2\ \rho \sum_{i,l,l^*} \frac{(\beta_2^q)^2 (x_i^q - l)(x_i^q - l^*)\eta_i \eta_i^*}{\left(1+\eta_i\right)^2\left(1+\eta_i^*\right)^2},$$

$$M_{3(u+3)} = \sum_{l=0}^{t-1} \frac{(\beta_2^q)^2 \beta_3^q (x_u^q - l)\eta_u^2}{\left(1+\eta_u\right)^4} + 2\ \rho \sum_{l,l^*} \frac{(\beta_2^q)^2 \beta_3^q (x_u^q - l)\eta_u \eta_u^*}{\left(1+\eta_u\right)^2\left(1+\eta_u^*\right)^2},$$

$$M_{(u+3)(u+3)} = \sum_{l=0}^{t-1} \frac{(\beta_2^q)^2 (\beta_3^q)^2 \eta_u^2}{\left(1+\eta_u\right)^4} + 2\ \rho \sum_{l,l^*} \frac{(\beta_2^q)^2 (\beta_3^q)^2 \eta_u \eta_u^*}{\left(1+\eta_u\right)^2\left(1+\eta_u^*\right)^2}, M_{ab} = 0, M_{vw} = M_{wv}.$$

In the simulation studies in 2.4, we assume independence of the measurements within replicate and then the terms involved with ρ are zero.

APPENDIX B

Now we provide the proof of Lemma 1 when we use the trace minimization procedure in Step 3 of the proposed algorithm. Lemma 1 also holds for component-wise minimization, which can be shown in a similar way. Suppose that the following conditions are satisfied:

(A1) $\frac{1}{ns/k}\sum_{q=1}^{s/k}\sum_{j=1}^{r}\sum_{l=0}^{t-1}\Sigma_{jl}^{q}\rightarrow\Sigma$ as $n\rightarrow\infty$ and $s\rightarrow\infty$,

(A2)

$$\frac{1}{(ns/k)^{3/2}}\sum_{q=1}^{s/k}\sum_{j=1}^{r}\sum_{l=0}^{t-1}E||\underline{S}_{jl}^{q}||^{3}\rightarrow 0 \text{ as } n\rightarrow\infty \text{ and } s\rightarrow\infty,$$

(A3) $(s/k)\leq n^{A}$ for some A,

where Σ_{jl}^{q}, Σ, and $\underline{S}_{jl}^{q}$ are defined in the proof.

Proof of Lemma 1:
In the paper we show that

$$\begin{aligned}\underline{\hat{\beta}}^{(c)} &= \sum_{q=1}^{s/k}V^{q}\underline{\hat{\beta}}^{q}\\ &= \underline{\beta}_{o}+\frac{1}{n}\sum_{q=1}^{s/k}\sum_{j=1}^{r}\sum_{l=0}^{t-1}V^{q}\,T\,\varphi_{IF}(y_{1jl}^{q},...,y_{kjl}^{q},\,\underline{\theta}_{o}^{q})\\ &+ \sum_{q=1}^{s/k}V^{q}\,T\,\underline{R}_{n}^{q},\end{aligned}$$

where T is a $3\times(3+k)$ matrix defined in (5) of the paper and

$$\varphi_{IF}(y_{1jl}^{q},...,y_{kjl}^{q},\underline{\theta}_{o}^{q})=\left(-\frac{1}{n}\sum_{j=1}^{r}\sum_{l=0}^{t-1}\dot{\psi}(y_{1jl}^{q},...,y_{kjl}^{q},\underline{\theta}_{o}^{q})\right)^{-1}\psi(y_{1jl}^{q},...,y_{kjl}^{q},\underline{\theta}_{o}^{q}).$$

It follows

$$\begin{aligned}\sqrt{ns/k}\,(\hat{\underline{\beta}}^{(c)} - \underline{\beta}_o) &= \frac{1}{\sqrt{ns/k}} \sum_{q=1}^{s/k} \sum_{j=1}^{r} \sum_{l=0}^{t-1} P^q\, T\, \varphi_{IF}(y^q_{1jl}, ..., y^q_{kjl}, \underline{\theta}^q_o) \\ &+ \sqrt{ns/k} \sum_{q=1}^{s/k} V^q\, T\, \underline{R}^q_n,\end{aligned}$$

where $P^q = (\frac{1}{s/k}\ \sum_{q=1}^{s/k} \Omega_q^{-1})^{-1}\ \Omega_q^{-1}$. Define a new random vector, $\underline{S}^q_{jl} = P^q\, T\, \varphi_{IF}(y^q_{1jl}, ..., y^q_{kjl}, \underline{\theta}^q_o)$. Denote $\Sigma^q_{jl} = var(\underline{S}^q_{jl})$.
Suppose that we have:

(A2′) For all $\epsilon > 0$,

$$\frac{1}{ns/k} \sum_{q=1}^{s/k} \sum_{j=1}^{r} \sum_{l=0}^{t-1} \int_{||\underline{s}|| > \epsilon\sqrt{ns/k}} ||\underline{s}||^2\, dF_i(\underline{s}) \ \to\ 0 \ \text{ as } \ n \to \infty \ \text{ and } \ s \to \infty.$$

By [21], under conditions **(A1)** and **(A2′)**, we can apply the Central Limit Theorem for independent non-identically distributed random vectors and thus

$$\frac{1}{\sqrt{ns/k}} \sum_{q=1}^{s/k} \sum_{j=1}^{r} \sum_{l=0}^{t-1} \underline{S}^q_{jl} \xrightarrow{d} N(\underline{0},\ \Sigma) \ \text{ as } \ n \to \infty \ \text{ and } \ s \to \infty. \qquad \text{(B.1)}$$

It is clear that $(A2')$ holds if we have $(A2)$, because

$$\begin{aligned}&\frac{1}{ns/k} \sum_{q=1}^{s/k} \sum_{j=1}^{r} \sum_{l=0}^{t-1} \int_{||\underline{s}|| > \epsilon\sqrt{ns/k}} ||\underline{s}||^2\, dF_i(\underline{s}) \\ =\ &\frac{1}{ns/k} \sum_{q=1}^{s/k} \sum_{j=1}^{r} \sum_{l=0}^{t-1} E\Big(||\underline{S}^q_{jl}||^2\, I\big(||\underline{S}^q_{jl}|| \geq \epsilon\sqrt{ns/k}\big)\Big) \\ \leq\ &\sum_{q=1}^{s/k} \sum_{j=1}^{r} \sum_{l=0}^{t-1} E\Big(\frac{||\underline{S}^q_{jl}||^2}{ns/k}\ \frac{||\underline{S}^q_{jl}||}{\epsilon\sqrt{ns/k}}\Big) \\ \leq\ &\frac{1}{\epsilon} \frac{\sum_{q=1}^{s/k} \sum_{j=1}^{r} \sum_{l=0}^{t-1} E||\underline{S}^q_{jl}||^3}{(ns/k)^{3/2}}.\end{aligned}$$

By [11], under $(A3)$, $||\underline{R}^q_n||$ is uniformly bounded in q such that

$$\max_q\ ||\underline{R}^q_n|| = O_p\Big(\frac{\big(log(s/k)\big)^2}{n}\Big),$$

and

$$\sqrt{ns/k}\sum_{q=1}^{s/k} V^q\, T\, \underline{R}_n^q \;\xrightarrow{p}\; \underline{0} \;\text{ as }\; n\to\infty \;\text{ and }\; s\to\infty. \tag{B.2}$$

Therefore, by Slutsky's Theorem,

$$||\underline{\hat{\beta}}^{(c)} - \underline{\beta}_o|| = O_p\big((ns)^{-1/2}\big) \;\text{ as }\; n\to\infty \;\text{ and }\; s\to\infty.$$

APPENDIX C

Now we prove Theorem 1 when we use the trace minimization procedure in Step 3 of the proposed algorithm. Theorem 1 also holds for component-wise minimization. Suppose that the following conditions are satisfied:

(B1) $\varphi_{IC}(\underline{\hat{\beta}}^{(c)})$ has continuous first and second derivatives in an open neighborhood of $\underline{\beta}_o$.

Proof of Theorem 1:
We use the same definition of $\varphi_{IC}(y_{ijl},\, \underline{\hat{\beta}}^{(c)}, x_{io})$ as in the paper,

$$\varphi_{IC}(y_{ijl},\, \underline{\hat{\beta}}^{(c)}, x_{io}) = \Big(-\frac{1}{n}\sum_{j=1}^{r}\sum_{l=0}^{t-1} \dot{\phi}(y_{ijl},\, \underline{\hat{\beta}}^{(c)}, x_{io})\Big)^{-1} \phi(y_{ijl},\, \underline{\hat{\beta}}^{(c)},\, x_{io}).$$

Denote the first and second derivatives of φ_{IC} evaluated at $\underline{\hat{\beta}}^{(c)} = \underline{\beta}$ by $(\varphi_{IC})^{(1)}_{\underline{\beta}}$ and $(\varphi_{IC})^{(2)}_{\underline{\beta}}$, respectively. Then the multivariate Taylor expansion ([22] and [23]) of φ_{IC} around $\underline{\beta}_o$ yields

$$\sqrt{n}(\tilde{x}_i - x_{io}) = \frac{1}{\sqrt{n}}\sum_{j=1}^{r}\sum_{l=0}^{t-1}\Big[\varphi_{IC}(y_{ijl}, \underline{\beta}_o, x_{io}) + [(\varphi_{IC})^{(1)}_{\underline{\beta}_o}]'\big(\underline{\hat{\beta}}^{(c)} - \underline{\beta}_o\big) + o_p\big(||\underline{\hat{\beta}}^{(c)} - \underline{\beta}_o||\big)\Big] + o_p(1).$$

It follows that

$$\begin{aligned}
\sqrt{n}\,(\tilde{x}_i - x_{io}) &= \frac{1}{\sqrt{n}}\sum_{j=1}^{r}\sum_{l=0}^{t-1} \varphi_{IC}(y_{ijl},\, \underline{\beta}_o,\, x_{io}) \\
&+ \frac{1}{\sqrt{n}}\sum_{j=1}^{r}\sum_{l=0}^{t-1} [(\varphi_{IC})^{(1)}_{\underline{\beta}_o}]'\ \frac{1}{n}\sum_{q=1}^{s/k}\sum_{j=1}^{r}\sum_{l=0}^{t-1} V^q\, T\, \varphi_{IF}(y^q_{1jl},\, \cdots, y^q_{kjl},\, \underline{\theta}^q_o) \\
&+ \frac{1}{\sqrt{n}}\sum_{j=1}^{r}\sum_{l=0}^{t-1} [(\varphi_{IC})^{(1)}_{\underline{\beta}_o}]'\ \sum_{q=1}^{s/k} V^q\, T\, \underline{R}^q_n \\
&+ \frac{1}{\sqrt{n}}\sum_{j=1}^{r}\sum_{l=0}^{t-1} o_p\big(||\underline{\hat{\beta}}^{(c)} - \underline{\beta}_o||\big) \\
&+ o_p(1). \qquad \text{(C.1)}
\end{aligned}$$

Let us now investigate the asymptotic behavior of each term on the right side of (C.1). The first term can be rewritten as follows and by the Central Limit Theorem it has an asymptotic normal distribution:

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{j=1}^{r} \sum_{l=0}^{t-1} \Big(- \frac{1}{n} \sum_{j=1}^{r} \sum_{l=0}^{t-1} \dot{\phi}(y_{ijl},\, \underline{\hat{\beta}}_{o}, x_{io}) \Big)^{-1} \phi(y_{ijl},\, \underline{\hat{\beta}}_{o},\, x_{io}) \\ \xrightarrow{d} \quad & N\Big(0,\, s_n^2 \Big) \ \text{ as } \ n \to \infty, \end{aligned}$$

where

$$s_n^2 = t\sigma^2 \Bigg[\sum_{l=0}^{t-1} \frac{(\beta_{2o})^2 (\beta_{3o})^2 e^{-2\beta_{3o}(x_{io}-l)}}{\big(1 + e^{-\beta_{3o}(x_{io}-l)}\big)^4} \Bigg]^{-1}.$$

The second term of (C.1) is

$$\frac{1}{\sqrt{s/k}} \ \frac{1}{n} \sum_{j=1}^{r} \sum_{l=0}^{t-1} [(\varphi_{IC})^{(1)}_{\underline{\beta}_o}]' \ \frac{1}{\sqrt{ns/k}} \sum_{q=1}^{s/k} \sum_{j=1}^{r} \sum_{l=0}^{t-1} \underline{S}^q_{jl}.$$

By the Weak Law of Large Numbers, it follows that

$$\frac{1}{n} \sum_{j=1}^{r} \sum_{l=0}^{t-1} [(\varphi_{IC})^{(1)}_{\underline{\beta}_o}]' \ \xrightarrow{p} \ E\Bigg[\frac{1}{t} \sum_{l=0}^{t-1} [(\varphi_{IC})^{(1)}_{\underline{\beta}_o}]' \Bigg] \ \text{ as } \ n \to \infty. \tag{C.2}$$

By (B.1), the second term goes to zero as $n \to \infty$ and $s \to \infty$. The third term of (C.1) is

$$\frac{1}{\sqrt{s/k}} \ \frac{1}{n} \sum_{j=1}^{r} \sum_{l=0}^{t-1} [(\varphi_{IC})^{(1)}_{\underline{\beta}_o}]' \ \sqrt{ns/k} \ \sum_{q=1}^{s/k} V^q \, T \, \underline{R}^q_n,$$

and it goes to zero as $n \to \infty$ and $s \to \infty$ by (B.2) and (C.2). The fourth term of (C.1) also converges to zero as $n \to \infty$ and $s \to \infty$ by the result of Lemma 1. Therefore,

$$\sqrt{n}\,(\tilde{x}_i - x_{io}) \xrightarrow{d} \ N\Big(0,\, s_n^2 \Big) \ \text{ as } \ n \to \infty \text{ and } s \to \infty,$$

where $s_n^2 = t\sigma^2 \Bigg[\sum_{l=0}^{t-1} \frac{(\beta_{2o})^2 (\beta_{3o})^2 e^{-2\beta_{3o}(x_{io}-l)}}{\big(1+e^{-\beta_{3o}(x_{io}-l)}\big)^4} \Bigg]^{-1}$.

REFERENCES

[1] M. F. Templin, D. Stoll, J. M. Schwenk, O. Potz, S. Kramer, and T. O. Joos, "Protein microarrays: promising tools for proteomic research," *Proteomics*, vol. 3, pp. 2155–2166, 2003.

[2] S. Nishizuka, L. Charboneau, L. Young, S. Major, W. C. Reinhold, M. Waltham, H. Kouros-Mehr, K. J. Bussey, J. K. Lee, V. Espina, P. J. Munson, E. Petricoin, L. A. Liotta, and J. N. Weinstein, "Proteomic profiling of the nci-60 cancer cell lines using new high-density reverse-phase lysate microarrays," *Proceedings of the National Academy of Sciences*, vol. 100, pp. 14 229–14 234, 2003.

[3] M. Sevecka and G. MacBeath, "State-based discovery: a multidimensional screen for small-molecule modulators of egf signaling," *Nature Methods*, vol. 3, pp. 825–831, 2006.

[4] B. Spurrier, P. Honkanen, A. Holway, K. Kumamoto, M. Terashima, S. Takenoshita, G. Wakabayashi, J. Austin, and S. Nishizuka, "Protein and lysate array technologies in cancer research," *Biotechnology Advances*, vol. 26, pp. 361–369, 2008.

[5] S. C. Wong, C. M. Chan, B. B. Ma, M. Y. Lam, G. C. Choi, T. C. Au, A. S. Chan, and C. A. T., "Advanced proteomic technologies for cancer biomarker discovery," *Expert review of proteomics*, vol. 6(2), pp. 123–134, 2009.

[6] M. Akkiprik, D. Nicorici, D. Cogdell, Y. J. Jia, A. Hategan, I. Tabus, O. Y. Yli-Harja, A. Sahin, and W. Zhang, "Dissection of signaling pathways in fourteen breast cancer cell lines using reverse-phase protein lysate microarray," *Technology in cancer research and treatment*, vol. 5, pp. 543–551, 2006.

[7] K. N. Mendes, D. Nicorici, D. Cogdell, I. Tabus, O. Yli-Harja, R. Guerra, S. R. Hamilton, and W. Zhang, "Analysis of signaling pathways in 90 cancer cell lines by protein lysate array," *Journal of Proteome Research*, vol. 6(7), pp. 2753–2767, 2007.

[8] S. Ramalingam, P. Honkanen, L. Young, T. Shimura, J. Austin, P. S. Steeg, and S. Nishizuka, "Quantitative assessment of the p53-mdm2 feedback loop using protein lysate microarrays," *Cancer Research*, vol. 67, pp. 6247–6252, 2007.

[9] F. Pirnia, M. Pawlak, G. G. Thallinger, B. Gierke, M. F. Templin, A. Kappeler, D. C. Betticher, B. Gloor, and M. M. Borner, "Novel functional profiling approach combining reverse phase protein microarrays and human 3d ex vivo tissue cultures: Expression of apoptosis related proteins in human colon cancer," *Proteomics*, vol. 9, pp. 3535–3548, 2009.

[10] I. Tabus, A. Hategan, C. Mircean, J. Rissanen, I. Shmulevich, Z. Wei, and J. Astola, "Nonlinear modeling of protein expressions in protein arrays," *IEEE transactions on signal processing*, vol. 54, pp. 2394–2407, 2006.

[11] X. He and Q. Shao, "A general bahadur representation of m-estimators and its application to linear regression with nonstochastic designs," *Annals of statistics*, vol. 24, pp. 2608–2630, 1996.

[12] C. Mircean, I. Shmulevich, D. Cogdell, W. Choi, Y. Jia, I. Tabus, S. R. Hamilton, and W. Zhang, "Robust estimation of protein expression ratios with lysate microarray technology," *Bioinformatics*, vol. 21, pp. 1935–1942, 2005.

[13] J. Hu, X. He, K. A. Baggerly, K. R. Coombes, B. T. Hennessy, and G. B. Mills, "Nonparametric quantification of protein lysate arrays," *Bioinformatics*, vol. 23, pp. 1986–1994, 2007.

[14] X. He and Q. Shao, "On parameters of increasing dimensions," *Journal of multivariate analysis*, vol. 73, pp. 120–135, 2000.

[15] E. S. Neeley, S. M. Kornblau, K. R. Coombes, and K. A. Baggerly, "Variable slope normalization of reverse phase protein arrays," *Bioinformatics*, vol. 25, pp. 1384–1389, 2009.

[16] L. Zhang, Q. Wei, L. Mao, W. Liu, G. B. Mills, and K. Coombes, "Serial dilution curve: a new method for analysis of reverse phase protein array data," *Bioinformatics*, vol. 25, pp. 650–654, 2009.

[17] D. Pena, "Combining information in statistical modeling," *The American Statistician*, vol. 51, pp. 326–332, 1997.

[18] E. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York: Wiley, 1983.

[19] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, 1st ed. Boca Raton, FL: Chapman and Hall/CRC, 1993.

[20] R. A. Johnson and D. Wichern, *Applied Multivariate Statistical Analysis*, 5th ed. Upper Saddle River, NJ: Prentice-Hall, 2002.

[21] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*, 1st ed. New York: Wiley, 1980.

[22] E. Lehmann, *Elements of Large-Sample Theory*, 1st ed. New York: Springer-Verlag, 1999.

[23] S. I. Resnick, *A Probability Path*, 1st ed. Boston: Birkhäuser, 1999.